



Bernhard Baltes-Götz

Statistisches Praktikum mit SPSS/PASW 17 für Windows

2010 (Rev. 100528)

Herausgeber:	Universitäts-Rechenzentrum Trier
	Universitätsring 15
	D-54286 Trier
	WWW: http://www.uni-trier.de/index.php?id=518
	E-Mail: urt@uni-trier.de
	Tel.: (0651) 201-3417, Fax.: (0651) 3921
Autor:	Bernhard Baltes-Götz (E-Mail: baltes@uni-trier.de)

Vorwort

Das seit Jahrzehnten bewährte und ständig aktualisierte Statistikprogramm **SPSS** (***S**tatistical **P**ackage for the **S**ocial **S**ciences*) wurde im Jahr 2009 umbenannt in **PASW** (***P**redictive **A**nalytics **S**oftware*), um wenige Monate später (nach Übernahme des Herstellers durch die Firma IBM) wieder den alten Namen zu erhalten. Genau genommen lautet der aktuelle Name **IBM SPSS Statistics**, während sich die im Manuskript beschriebene Programmversion 17 als **SPSS Statistics** bezeichnet. Wir verwenden im Manuskript den kompakten Namen **SPSS** und reden dabei über ein weitgehend komplettes und relativ leicht zu bedienendes Statistikprogramm, das in den Geo-, Wirtschafts- und Sozialwissenschaften sehr verbreitet ist und alle wichtigen Betriebssysteme für Arbeitsplatzrechner unterstützt (Linux, MacOS, Windows).

Im vorliegenden Manuskript wird ein Einblick in die statistische Datenanalyse mit der SPSS-Version 17 für Windows vermittelt, wobei großer Wert auf die methodologische Einordnung der beschriebenen EDV-Techniken gelegt wird. Wesentliche Teile des Manuskripts sind wegen der weitgehend konsistenten Bedienungslogik auch für andere SPSS-Versionen unter Windows oder alternativen Betriebsystemen verwendbar.

Dieses Manuskript dient primär als Begleitlektüre zum Kurs *Statistisches Praktikum mit SPSS für Windows* am Universitäts-Rechenzentrum Trier (URT) erstellt, kann jedoch auch im Selbststudium verwendet werden. Dass dabei die meisten Themen in konkreter Arbeit am Rechner nachvollzogen werden sollten, folgt aus der Kurskonzeption:

Zielgruppe/Voraussetzungen

Der Kurs ist konzipiert für Personen, die in wesentlichem Umfang bei Forschungsarbeiten mit SPSS mitwirken wollen, also z.B. im Rahmen einer Abschlussarbeit die Durchführung einer eigenen Studie planen oder bereits begonnen haben. Wer lediglich einfache Teilaufgaben zu erledigen hat (z.B. wenige Auswertungen mit einer bereits vorhandenen und fehlerbereinigten SPSS-Datendatei), der sollte eventuell die zweistündige SPSS-Kurzeinführung des Rechenzentrums besuchen oder das zugehörige Manuskript lesen.[1]

Im Kurs wird eine methodische Grundausbildung (empirische Forschung, Statistik) vorausgesetzt, wie sie üblicherweise in den Studiengängen empirisch-statistisch forschender Disziplinen vermittelt wird.

An EDV-Voraussetzungen werden nur elementare Fertigkeiten im Umgang mit PCs unter Windows erwartet.

Kursinhalte

Wir konzentrieren uns darauf, in anderen Veranstaltungen (z.B. zur empirischen Forschung oder Statistik) erlernte Begriffe und Methoden mit dem EDV-Werkzeug SPSS in der Praxis anzuwenden. Zwar werden im Kursverlauf viele methodische Themen in knapper Form behandelt, doch kann damit eher vorhandenes Wissen aufgefrischt als neues erworben werden. Insbesondere kann die Anwendung und Diskussion der vielfältigen statistischen Auswertungsmethoden nur exemplarisch stattfinden. Eine explizite Behandlung ist nur bei wenigen, besonders häufig eingesetzten Verfahren möglich (z.B. bei der Kreuztabellenanalyse). Zu zahlreichen Auswertungsmethoden bietet das Rechenzentrum Spezialveranstaltungen an, in denen die wesentlichen methodologischen Grundlagen und natürlich die praktische Durchführung mit SPSS erläutert werden.

[1] Es ist als PDF-Dokument auf dem Webserver der Universität Trier an der selben Stelle zu finden wie die aktuelle Ausgabe des vorliegenden Manuskripts (siehe unten).

Informationen über das URT-Kursprogramm finden Sie z.B. auf dem WWW-Server der Universität Trier von der Startseite (http://www.uni-trier.de/) ausgehend über:

Rechenzentrum > Infos für Studierende > Kursangebot

Zu den meisten Kursen sind ausführliche Manuskripte entstanden, die Sie auf dem Webserver der Universität Trier folgendermaßen finden:

Rechenzentrum > Infos für Studierende > EDV-Dokumentationen > Statistik

Im Sinne einer praxisnahen, projektorientierten Ausbildung beschreibt das Manuskript eine vollständige empirische Studie von der ersten Idee über die Kodierung, Erfassung, Kontrolle und Modifikation der Daten bis zur statistischen Auswertung und zur Verwertung der Ergebnisse.

Zwar werden auch in EDV-handwerklicher Sicht die SPSS-Optionen nicht annähernd vollständig behandelt, doch sollten Sie nach dem Kurs mit den erworbenen Grundkenntnissen unter Verwendung der aufgezeigten Informationsmöglichkeiten selbständig und erfolgreich mit SPSS arbeiten können.

Zugriff auf die Dateien zum Kurs

Die aktuelle Version des Manuskripts ist als PDF-Dokument zusammen mit den im Kurs benutzen Dateien auf dem Webserver der Universität Trier von der Startseite (http://www.uni-trier.de/) ausgehend folgendermaßen zu finden:

Rechenzentrum > Studierende > EDV-Dokumentationen >
Statistik > Statistisches Praktikum mit SPSS für Windows

Leser(innen) im Selbststudium werden in der Regel keine eigene Datenerhebung realisieren, können jedoch mit den zur Verfügung gestellten Dateien alle Projektarbeitsschritte ab der Datenprüfung konkret durchführen.

Kritik und Verbesserungsvorschläge zum Manuskript werden dankbar entgegen genommen (z.B. unter der Mail-Adresse baltes@uni-trier.de).

Trier, im Mai 2010 Bernhard Baltes-Götz

Inhaltsverzeichnis

1 Von der Theorie zu den SPSS-Variablen

1.1 Statistik und EDV als Hilfsmittel der Forschung

Die Erfahrungswissenschaften bemühen sich um allgemeingültige Aussagen deskriptiver, explanatorischer oder prognostischer Art. In vielen Anwendungsbereichen sind dabei *deterministische* Gesetze (z.B. Ohmsches Gesetz der Elektrik, Hebelgesetz der Mechanik) kaum zu finden, und man benötigt eine Technologie zur Untersuchung *probabilistischer* Gesetze.

Beispiel: Welchen Effekt hat das Rauchen auf die Entstehung von Lungenkrebs?

Wie wir wissen, hat das Rauchen auch bei gleicher Dosierung der Schadstoffe keinesfalls für alle Personen dieselben Folgen. Die Frage nach dem Effekt des Rauchens ist anhand weniger, unrepräsentativer Einzelbeobachtungen (z.B. der steinalte Kettenraucher) nicht zu klären. In einer solchen Situation können statistische Methoden dabei helfen, rationale Entscheidungen zu treffen, denn Wallis & Roberts (1956, S. 1) stellen treffen fest:

> "Statistics is a body of methods for making wise decisions in the face of uncertainty."

Eine grundlegende Strategie der statistisch arbeitenden Forschung, trotz Unsicherheit zu guten Entscheidungen zu kommen, besteht darin, zu einer Fragestellung hinreichend viele **unabhängige** Beobachtungen zu machen, um aus dieser **Stichprobe** Informationen über die zugrunde liegende **Population** der potentiellen Beobachtungen zu gewinnen. Zur Untersuchung der Raucherproblematik wird man vielleicht 100 Raucher und 100 Nichtraucher (= **Beobachtungseinheiten, Merkmalsträger, Fälle**) auf das Vorliegen einer Lungenkrebserkrankung untersuchen, so dass die beiden **Merkmale** *Raucher* und *Lungenkrebs* (jeweils mit den Ausprägungen *Ja* und *Nein*) resultieren, z.B. mit der folgenden Stichprobenverteilung:

		Lungenkrebs		
		Ja	Nein	Gesamt
Raucher	Ja	30	70	100
	Nein	1	99	100
	Gesamt	31	169	200

Mit statistischen Forschungsmethoden lassen sich u.a. folgende Fragestellungen bearbeiten:

- **Parameterschätzung**

  Beispiel: Wie groß ist die Wahrscheinlichkeit, an Lungenkrebs zu erkranken, bei Rauchern bzw. Nichtrauchern?

  Um z.B. die Lungenkrebs-Wahrscheinlichkeit für Raucher (also einen Populationsparameter) anhand von Stichprobendaten zu schätzen, verwendet man die relative Häufigkeit von Lungenkrebs in der Raucherteilstichprobe (= 0,3). In der Regel wird man sich nicht auf **Punktschätzungen** beschränken, sondern zusätzlich auch **Intervallschätzungen** vornehmen. Dabei wird aus den Stichprobendaten für jeden fraglichen Populationsparameter ein **Vertrauensintervall** (synonym: **Konfidenzintervall**) ermittelt, das seinen wahren Wert mit einer gewünschten Wahrscheinlichkeit (z.B.: 0,95) enthält.

- **Hypothesentests** (konfirmatorische Verfahren)

  Beispiel: Ist bei Rauchern das Lungenkrebsrisiko größer als bei Nichtrauchern?

  Hier ist eine Entscheidung zwischen *zwei* Hypothesen zu treffen:

  - **Nullhypothese**

    Im Beispiel: Das Lungenkrebsrisiko ist bei Rauchern *nicht* größer als bei Nichtrauchern.

  - **Alternativhypothese**

    Im Beispiel: Das Lungenkrebsrisiko ist bei Rauchern erhöht.

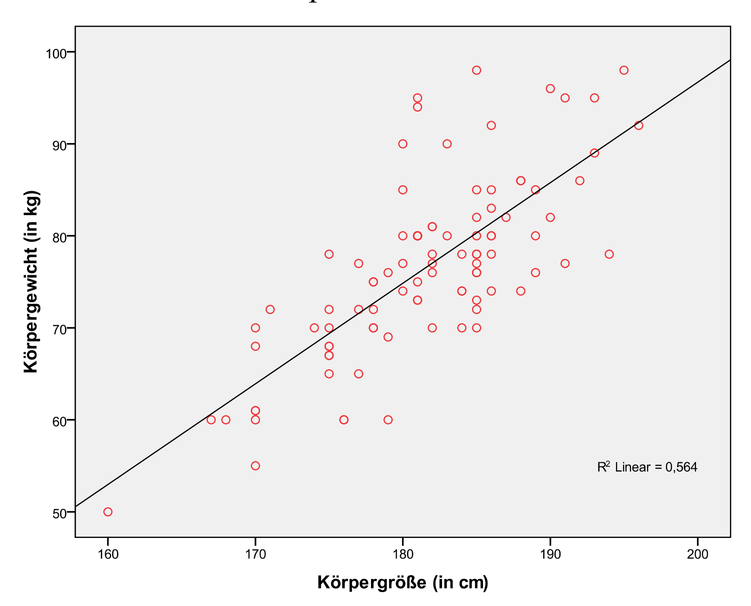
Auch zur Klärung der Frage, wie viele Beobachtungen erforderlich sind, um einen Effekt bestimmter Größe (im Beispiel: eine Risikodifferenz) mit einer bestimmten Wahrscheinlichkeit nachweisen zu können, stehen statistische Methoden bereit (siehe unten).

- **Modellierung**

  Die zu schätzenden bzw. auf Signifikanz zu prüfenden Parameter stammen aus einem mathematischen Modell, das auf ein empirisches System angewendet wird. Im Raucherbeispiel kommt ein extrem einfaches Modell mit binomialverteilten (dichotomen) Zufallsvariablen zu Einsatz, das kaum als Modell in Erscheinung tritt. Daher betrachten wir an dieser Stelle noch das häufig verwendete Modell der linearen Regression, das in seiner einfachsten Form ein zu erklärende Kriterium Y, einen Regressor X und ein Residuum ε enthält:

  $$Y = \beta_0 + \beta_1 X + \varepsilon$$

  Es eignet sich z.B. dazu, den Einfluss der Körpergröße auf das Körpergewicht von Personen zu modellieren. Im folgenden Streudiagramm ist die gemeinsame empirische Verteilung der beiden Merkmale in einer Stichprobe zu sehen:



  Offenbar ist im Beispiel die von den beiden Modellparametern β_0 und β_1 abhängige Regressionsgerade gut geeignet, den erwarteten Y-Wert für eine gegebene X-Ausprägung vorherzusagen. Modellgültigkeit vorausgesetzt, sind das Konfidenzintervall und der Signifikanztest zum Steigungsparameter β_1 von zentralem Interesse. Bei den meisten Fragestellungen sind *mehrere* Einflussgrößen zu untersuchen und z.B. in ein multiples Regressionsmodell aufzunehmen.

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

  Ein Forschungs*programm* wird nicht bei der Untersuchung *eines* Kriteriums stehen bleiben, sondern nach einem Modell zur Beschreibung und Erklärung eines empirischen Systems suchen. Man kann z.B. versuchen, das Zusammenwirken aller relevanten Komponenten durch ein Pfad- oder ein Strukturgleichungsmodell zu erfassen. Hier sind zahlreiche Parameterschätzungen und Hypothesentests beteiligt.

Kehren wir kurz zurück zum medizinischen Beispiel. In einer realen Studie wird man sich nicht auf die oben zur Illustration verwendeten dichotomen Merkmale beschränken, sondern Dauer und Ausmaß des (aktiven und passiven) Rauchens sowie den gesundheitlichen Status der Probanden genauer untersuchen und außerdem viele zusätzliche Merkmale erheben, z.B. Alter, Geschlecht, Beruf, Schadstoffbelastung der Wohnung (speziell Radon). Eine praktikable Auswertung solcher Datenmengen ist nur mit EDV-Hilfe möglich. Mit SPSS für Windows steht ein be-

quemes, leistungsfähiges und sehr bewährtes Analysesystem für die statistische Forschung zur Verfügung. Es bietet u.a. …

- fast alle wichtigen statistischen Verfahren
- gute graphische Darstellungsmöglichkeiten
- eine umfangreiche Unterstützung bei der Datenverwaltung
- die in der Windows-Welt gebräuchlichen Verfahren zur Kooperation mit anderen Programmen (z.B. Zwischenablage, Programmierschnittstellen)

Weil SPSS auch auf anderen Plattformen vertreten ist und sein Datendateiformat weithin unterstützt wird, bestehen günstige Bedingungen für die kollegiale Kommunikation.

1.2 Planung und Durchführung einer empirischen Untersuchung im Überblick

Zunächst wollen wir uns einen Überblick über die verschiedenen Phasen eines empirischen Forschungsprojekts und damit auch über unser Kursprogramm verschaffen. Dabei werden zahlreiche Aufgaben, Methoden und Probleme angesprochen, über die Sie sich im Bedarfsfall in den Lehrveranstaltungen oder in der Literatur zur empirischen Forschung informieren können (siehe z.B. Bortz & Döring 1995; Pedhazur & Pedhazur Schmelkin 1991; Schnell, Hill & Esser 2005). Die anschließende Darstellung soll als Übersicht dienen und ist daher relativ knapp gehalten. Ihr folgt unmittelbar die konkrete und ausführliche Anwendung auf unsere Beispielstudie. Weil die dargestellten Aufgaben teilweise interdependent sind, bilden sie keine strenge, bei allen empirischen Studien gleichförmig ablaufende Sequenz.

1.2.1 Forschungsziele, Hypothesen und Modelle

Einer empirischen Untersuchung wird in der Regel eine längere Phase der intensiven theoretischen Auseinandersetzung mit dem Thema vorangehen. Daraus ergeben sich Forschungsinteressen, die - u.a. in Abhängigkeit vom Forschungsstand - eher von **explorativer** (hypothesensuchender) oder eher von **konfirmatorischer** (hypothesenprüfender) Natur sind. Oft werden *beide* Forschungsstrategien vertreten sein. Die zu prüfenden Hypothesen sollten wegen ihrer Steuerungsfunktion für spätere Schritte möglichst exakt formuliert werden. Häufig werden sich die Hypothesen auf Parameter in einem **mathematischen Modell** (z.B. in einer linearen Regressionsgleichung) beziehen.

1.2.2 Untersuchungsplanung

Wenn Sie eine Theorie bzw. eine Hypothesenfamilie empirisch prüfen oder einen Gegenstandsbereich empirisch explorieren möchten, haben Sie bei der Untersuchungsplanung zahlreiche Aufgaben zu lösen:

- **Festlegung der Beobachtungseinheit(en) und der zu untersuchenden Merkmale**
  In der Regel ergibt sich aus der Fragestellung unmittelbar, welche Beobachtungseinheiten (Merkmalsträger) einer Studie zugrunde liegen sollten (z.B. Personen, Volkswirtschaften, Orte, Betriebe, Bodenproben, Jahre), und welche Merkmale bei jeder Beobachtungseinheit festgestellt werden sollten.

  Gelegentlich bieten sich **hierarchisch geschachtelte Untersuchungseinheiten** auf mehreren Ebenen an (siehe z.B. Raudenbush & Bryk 2002). So hat man es etwa bei einer Studie zur Arbeitszufriedenheit und Produktivität von Arbeitnehmern aus verschiedenen Firmen in Abhängigkeit von Person- und Organisationsmerkmalen mit Beobachtungseinheiten auf zwei Ebenen zu tun:
    - Arbeitnehmer
    - Firmen

Bei der späteren Auswertung ist zu beachten, dass traditionelle statistische Methoden (z.B. die lineare Regressionsanalyse) *unabhängige Residuen* annehmen. Die bei einer hierarchischen Datenstruktur auf der untersten Ebene naturgemäß anzutreffende Abhängigkeit der *Beobachtungen* muss in speziellen Modellen berücksichtigt werden, um gültige Vertrauensintervalle und Hypothesentests zu erhalten. Das Demonstrationsprojekt in unserm Kurs kommt mit einer konventionellen, flachen Datenstruktur aus, und die Behandlung der speziellen Optionen und Probleme der Mehrebenenanalyse bleibt einem speziellen Kurs vorbehalten.

- **Entscheidung für ein Untersuchungsdesign**
  Sie können z.B. einen (quasi-)experimentellen Untersuchungsplan entwerfen oder eine reine Beobachtungsstudie wählen, die quer- oder längsschnittlich angelegt sein kann. Zur Prüfung einer Theorie ist eine empirische Situation zu wählen bzw. zu gestalten, die zum Anwendungsbereich der Theorie gehört.

- **Operationalisierung der zu untersuchenden Merkmale**
  Zur Operationalisierung von theoretischen Begriffen (z.B. sozioökonomischer Status, Ärger, Optimismus) sollten möglichst valide und reliable Messmethoden gewählt bzw. entworfen werden, die außerdem nicht zu aufwändig sind. Das Skalenniveau der Messmethoden muss die Voraussetzungen der geplanten statistischen Auswertungsverfahren (siehe unten) erfüllen.

  Bei *quantitativen* Merkmalen (z.B. Alter) sollten die verfügbaren Informationen bei der Erfassung *nicht* durch eine *künstliche* und *willkürliche* Klassenbildung reduziert werden (z.B. durch Bildung der Altersklassen < 20, 21- 40, 41-60, > 60). Häufig sind Modelle für metrische Daten einfacher und erfolgreicher als solche für vergröberte Daten. Vor allem kann man mit SPSS zu einer metrischen Variablen nach Belieben klassifizierte Varianten erzeugen, wenn dies für spezielle Analysen wünschenswert erscheint. Eine Ausnahme von der Empfehlung zur Erfassung metrischer Informationen ist z.B. bei der Befragung von Personen nach ihrem Einkommen zu machen. Um bei dieser sensiblen Frage Widerstände zu vermeiden, muss man sich in der Regel auf die Erhebung von groben Einkommensklassen beschränken.

  Bei den Überlegungen zur Operationalisierung spielen auch die verfügbaren technischen Hilfsmittel für die Datenerhebung und -erfassung eine Rolle. Mit Hilfe der Computer-Technik ist eine interaktive, individualisierte und dabei auch noch ökonomische Datenerfassung möglich. Bei der zeitgenauen Steuerung experimenteller Abläufe kommen spezielle Rechner im Forschungslabor zum Einsatz. Für eine kontinuierliche, alltagsbegleitende Datenerfassung können oft Rechner im Taschenformat genutzt werden. Einfache Befragungen werden mittlerweile routinemäßig via Internet realisiert, wenn die zu untersuchende Population auf diesem Weg erreichbar ist.

- **Empirisch prüfbare Hypothesen (über Modellparameter) formulieren**
  Aus einer in theoretischen Begriffen formulierten Hypothese ergibt sich im Verlauf der Untersuchungsplanung durch zahlreiche Konkretisierungen und Operationalisierungen eine in empirischen Begriffen formulierte und damit statistisch prüfbare Hypothese.

  Von einfachen Fällen abgesehen, werden sich die zu prüfenden Hypothesen einer Studie auf Parameter eines mathematischen Modells beziehen, das also spätestens jetzt explizit zu formulieren ist.

  Bei den Hypothesen muss klar erkennbar sein, ob eine *gerichtete* oder eine *ungerichtete* Behauptung vorliegt. Über den Steigungsparameter β_1 der bivariaten linearen Regression mit dem Kriterium Y und dem Regressor X:

  $$Y = \beta_0 + \beta_1 X + \varepsilon$$

  behauptet eine gerichtete Alternativhypothese z.B.

$$\beta_1 > 0$$

Dem steht die folgende Aussage der ungerichteten Hypothese gegenüber:

$$\beta_1 \neq 0$$

Sie ist schwächer, weil die Richtung der Abweichung vom neutralen Wert Null offen bleibt.

- **Statistische Versuchsplanung**
  Für jede Hypothese ist ein **statistisches Entscheidungsverfahren** zu wählen, dessen Voraussetzungen an Skalenniveau und Verteilungsverhalten der beteiligten Merkmale (voraussichtlich) erfüllt sind. Zu jedem Test ist das **Fehlerrisiko erster Art** (α-Fehler) festzulegen, wobei z.B. die 5%-Konvention übernommen werden kann.

  Es ist zu überlegen, wie eine repräsentative und zur Durchführung der geplanten Auswertungsverfahren hinreichend große **Stichprobe** rekrutiert werden kann. Bei ausgeprägt konfirmatorisch angelegten Studien ist bei der Stichprobenumfangsplanung insbesondere das **Fehlerrisiko zweiter Art** (der β-Fehler) zu berücksichtigen.

- **Strukturierung und Kodierung der Daten**
  Wer ganz sicher gehen will, dass die bei einer Studie erhobenen Informationen sicher und bequem in die EDV übernommen werden können, sollte die Daten schon in der Planungsphase gegenüber der zuständigen Software deklarieren. Beim Entwurf eines Formulars für eine Online-Erhebung oder für eine Datenerfassung per Scanner geschieht die Datendeklaration gegenüber der jeweils verwendeten Software (also vor der Datenerhebung). Diese Software kann in der Regel die erfassten Merkmale später als SPSS-Datendatei exportieren, so dass keine erneute Datendeklaration gegenüber SPSS erforderlich ist. Häufig werden die Daten mit schriftlichen Untersuchungsdokumenten erhoben und anschließend manuell erfasst. Man sollte auch bei diesem Vorgehen die Daten schon vor der Erhebung gegenüber dem geplanten Erfassungsprogramm (z.B. SPSS-Dateneditor) deklarieren. Anfänger(innen) werden bei der Arbeit mit einem Computer-Programm, das die vorwiegend forschungslogisch und kaum durch EDV-Restriktionen diktierte Datenstruktur explizit einfordert, konzeptionelle Probleme eventuell eher entdecken als bei der schriftlichen Beschreibung des Forschungsvorhabens.

  Bei den meisten Projekten können die Daten in *einer* Matrix (Tabelle) mit den Fällen als Zeilen und den Merkmalen als Spalten untergebracht werden. Gelegentlich werden *mehrere* Tabellen benötigt, z.B. bei einer Untersuchung von Mitarbeitern und Kunden einer Einzelhandelskette.

  Bei einer flachen Datenstruktur (ohne geschachtelte Beobachtungseinheiten, siehe oben) sind oft nur Kodierungsregeln festzulegen. Hierunter fällt z.B. die Vereinbarung, dass beim Merkmal Geschlecht die Ausprägung *weiblich* durch eine Eins und die Ausprägung *männlich* durch eine Zwei erfasst werden soll.

  Die Festlegungen zur Strukturierung und Kodierung der Projektdaten sollten in einem **Kodierplan** dokumentiert werden. Er ist bei einer manuellen Datenerfassung als genaue Arbeitsvorschrift unverzichtbar und eignet sich generell zur Dokumentation der Daten (eventuell für einen größeren Nutzerkreis).Wir werden uns in Abschnitt 1.4 mit der Strukturierung und Kodierung von Daten ausführlich beschäftigen.

1.2.3 Durchführung der Studie (inklusive Datenerhebung)

Nach Abschluss der Planungs- und Vorbereitungsphase kann die Studie durchgeführt werden.

1.2.4 Datenerfassung und -prüfung

Nach einer Datenerhebung per Fragebogen stehen als Nächstes folgende Arbeiten an:

- **Datenerfassung**
  Das Eintragen der Rohdaten in eine Datei auf der Festplatte eines Computers kann mit dem Dateneditor von SPSS geschehen oder mit einem speziellen Datenerfassungsprogramm. In jedem Fall ist bei der Erfassung der in der Planungsphase oder spätestens nach der Datenerhebung erstellte Kodierplan genau einzuhalten. Hier ist z.B. für jedes Merkmal festgelegt, wie seine Ausprägungen kodiert werden sollen.

  Bei schriftlichen Befragungen großer Stichproben kann eine Anlage zum automatischen Einscannen und Interpretieren von Untersuchungsdokumenten rentabel eingesetzt werden. Voraussetzung ist dann u.a. die Beachtung einiger Regeln beim Entwurf der Untersuchungsmaterialien.
- **Überprüfung auf Erfassungsfehler**
  Je fehleranfälliger die gewählte Erfassungsmethode war, desto mehr Aufwand ist bei der Datenprüfung angebracht.

Bei einer Online-Datenerhebung entfällt die Datenerfassung. Im Abschnitt 3.1.1 folgende weitere Informationen zu den Techniken der automatischen Datenerhebung- bzw. -erfassung.

1.2.5 Datentransformation

Nach der Erfassung und Prüfung liegen bei vielen Studien die Daten immer noch nicht in auswertbarer Form vor. Vielfach müssen Variablen überarbeitet (z.B. rekodiert) oder aus Vorläufern neu berechnet werden (z.B. durch Mittelwertsbildung). Solche Transformationen nehmen bei vielen Projekten einen erheblichen Umfang an, wobei sowohl akribische Fleißarbeit als auch kreative Begriffsbildung gefragt sind.

1.2.6 Statistische Datenanalyse

Nach langer Mühe können mit Hilfe von SPSS z.B. die gesuchten Schätzwerte (samt Konfidenzintervallen) ermittelt und die geplanten Hypothesentests durchgeführt werden. Bei einer eher explorativen Untersuchungsanlage ist eine längere, kreative Auseinandersetzung mit den Daten erforderlich, wobei zahlreiche Datentransformationen und statistische Analysen ausgeführt werden.

1.3 Beispiel für eine empirische Untersuchung

Um die im Rahmen einer empirischen Untersuchung mit SPSS zu erledigenden Arbeiten unter realistischen Bedingungen üben zu können, wird im Verlauf des Kurses eine kleine psychologische Fragebogenstudie durchgeführt.[1] Dabei werden Sie alle Phasen der empirischen Forschung von der ersten Idee bis zur statistischen Hypothesenprüfung mit SPSS kennen lernen und die erforderlichen Arbeiten zum großen Teil selbständig durchführen. Als Beispiel wurde u.a. deshalb eine psychologische Fragebogenstudie gewählt, weil die Kursteilnehmer dabei in wenigen Minuten interessante empirische Daten selbst erzeugen können. Damit ist auch die Phase der Daten*erhebung* in den Übungsablauf einbezogen, die ansonsten aus Zeitgründen ausgespart werden müsste.

---

[1] Hierbei werden in stark vereinfachter Form Ideen aus einem ehemaligen Forschungsprojekt von Herrn Prof. Dr. J. Brandtstädter (Universität Trier) aufgegriffen, dem ich an dieser Stelle herzlich für die Erlaubnis und für die Überlassung von Untersuchungsmaterial danken möchte.

Bezogen auf das in Abschnitt 1.2 vorgestellte Ablaufschema beschäftigen wir uns nun mit dem theoretischen Hintergrund unserer Studie und mit Fragen der Untersuchungsplanung.

1.3.1 Die allgemeinpsychologische KFA-Hypothese

Nach einer Theorie von Kahneman[1] & Miller (1986) hängt die Stärke unserer emotionalen Reaktion auf ein positives oder negatives Ereignis u.a. davon ab, welche alternativen (aber nicht eingetretenen) Ereignisse wir uns vorstellen können, mit anderen Worten: welche **kontrafaktischen Alternativen** mental verfügbar sind. Wir wollen uns auf den Fall ungünstiger Ereignisse beschränken. Hierfür stellen Kahneman & Miller die folgende Hypothese auf:

> **Bei einem negativen Ereignis erhöht die mentale Verfügbarkeit (Vorstellbarkeit) kontrafaktischer (also positiver) Ereignisalternativen den erlebten Ärger.**

Im weiteren Verlauf wollen wir unser Projekt kurz als *KFA-Studie* bezeichnen.

Weil diese Hypothese für beliebig aus der Population herausgegriffene Personen Gültigkeit beansprucht, kann sie als *allgemeinpsychologisch* bezeichnet und von *differentialpsychologischen* Hypothesen unterschieden werden, die sich mit Unterschieden zwischen Personen beschäftigen (siehe Abschnitt 1.3.3).

1.3.2 Untersuchungsplanung

Hinsichtlich des Untersuchungsdesigns haben wir uns aufgrund praktischer Erwägungen bereits auf eine **querschnittlich angelegte Fragebogenstudie** mit den Kursteilnehmern als **Beobachtungseinheiten** festgelegt. Zentrales **Merkmal** ist der Ärger über ein negatives Ereignis bei An- bzw. Abwesenheit einer kontrafaktischen (also positiven) Ereignisalternative.

Nun geht es um den Entwurf des Fragebogens, wobei z.B. der theoretische Begriff *Ärger* zu **operationalisieren** ist. Wir wollen die Untersuchungsteilnehmer bitten, sich in eine Geschichte einzufühlen, bei der zwei Personen objektiv denselben Schaden erleiden, jedoch in unterschiedlichem Grad eine kontrafaktische (also günstige) Alternative vor Augen haben. Dann sollen die Probanden für jeden Geschädigten angeben, wie stark sie sich in dessen Lage ärgern würden. Die genaue Instruktion ist dem unten wiedergegebenen Fragebogen (Teil 2) zu entnehmen. Die beiden Ärgermessungen werden durch Ratingskalen realisiert, wobei das Antwortformat der Anschaulichkeit halber an ein Thermometer mit den Ankerpunkten 0° und 100° erinnert. Wir gehen davon aus, dass die Ärgermessungen annähernd Intervallniveau besitzen.

Indem wir jede Person den *beiden* imaginierten Behandlungen aussetzen, gewinnen wir jeweils *zwei* Beobachtungswerte, die eine statistische Analyse der allgemeinpsychologischen Hypothese mit relativ hoher Teststärke (kleinem β-Fehler) ermöglichen sollen. Gegen diese Befragungstechnik lässt sich einwenden, dass durch die Präsentation der *beiden* Varianten ein Kontrast künstlich induziert, zumindest jedoch verstärkt wird. Um diese **Artefaktgefahr** zu vermeiden, könnte man statt des Messwiederholungsfaktors KFA einen Gruppierungsfaktor verwenden und jede Person nur zu *einer* Schädigungsvariante befragen. Weil das Artefaktargument nicht zwingend und die Kursstichprobe aus organisatorischen Gründen relativ klein ist, hat das Teststärkeargument ein höheres Gewicht.

In Abschnitt 1.3.1 wurde die KFA-Hypothese noch ohne Bezug auf unsere Untersuchungsplanung formuliert. Jetzt nehmen wir eine Konkretisierung vor durch …

[1] Kahneman erhielt 2002 den Nobelpreis für Wirtschaft, womit vor allem seine erfolgreiche Anwendung psychologischer Erkenntnisse (u.a. zu Urteilen und Entscheidungen unter Unsicherheit) in wirtschaftswissenschaftlichen Theorien gewürdigt wurde.

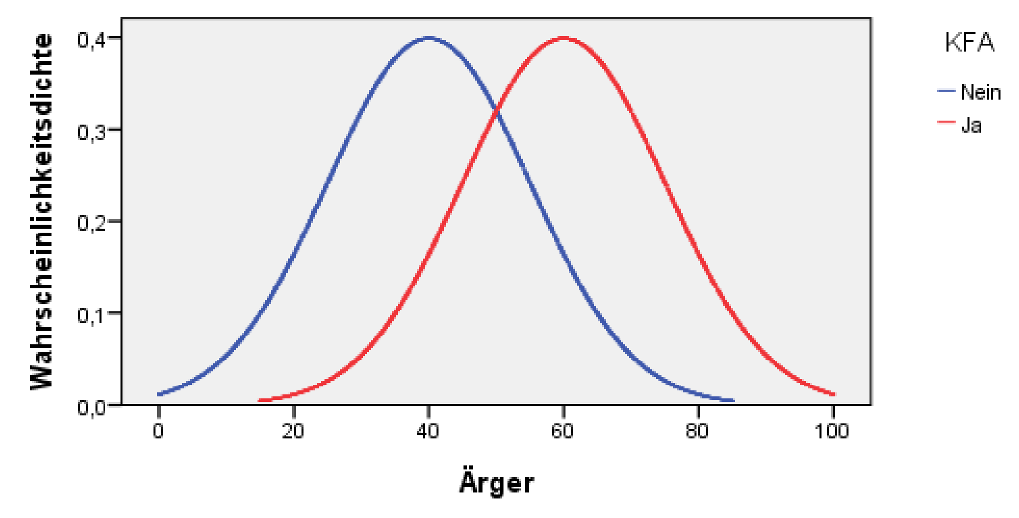
- Verwendung von direkt beobachtbaren Begriffen
- Bezug auf Verteilungsparameter (Erwartungs- bzw. Mittelwert)
  Eingangs wurde betont, dass unsere Hypothesen in der Regel probabilistischer Natur sind. Auch bei einer allgemeinpsychologischen Hypothese wird man kaum auf einer Gültigkeit für *alle* Personen einer Population bestehen (womöglich sogar mit derselben Effektstärke). Die konkretisierte Hypothese sollte über die im statistischen Entscheidungsverfahren tatsächlich analysierten Modell- bzw. Verteilungsparameter reden.

Außerdem soll hier der Klarheit halber (in einer für Forschungsberichte kaum zu empfehlenden Ausführlichkeit) dargelegt werden, dass bei einem inferenzstatistischen Entscheidungsverfahren *zwei* konkurrierende Hypothesen beteiligt sind:

Nullhypothese: **Die Untersuchungsteilnehmer erleben in der Rolle des Geschädigten mit hochgradig verfügbarer kontrafaktischer Alternative im Mittel *nicht* mehr Ärger als in der Rolle des Geschädigten mit "weit entfernter" kontrafaktischer Alternative.**

Alternativhypothese:[1] **Die Untersuchungsteilnehmer erleben in der Rolle des Geschädigten mit hochgradig verfügbarer kontrafaktischer Alternative im Mittel mehr Ärger.**

Das folgende Diagramm zeigt die Verteilungen des Ärgers aus der Situation *ohne* (blau) bzw. *mit* KFA (rot) im Sinne unserer Alternativhypothese:



Wir wollen unser Entscheidungsproblem mit einem **t-Test für verbundene** bzw. **abhängige Stichproben** lösen, falls die Verteilungsvoraussetzungen dieses Verfahrens erfüllt sind. Da gerichtete Hypothesen vorliegen, ist **einseitig** zu testen. Dabei wird eine Irrtumswahrscheinlichkeit erster Art in Höhe von **α = 5%** akzeptiert.

Unsere Studie soll aus praktischen Gründen mit der **studentischen Stichprobe** der Kursteilnehmer durchgeführt werden. Damit können unter induktivistischer Perspektive die Ergebnisse günstigstenfalls auf die Population der Studierenden generalisiert werden.

Da aus statistischer Sicht eine Stichprobe nie zu groß sein kann, sollen nach Möglichkeit *alle* Kursteilnehmer als Probanden gewonnen werden. Es ist aus praktischen Gründen nicht möglich, weitere Untersuchungsteilnehmer zu rekrutieren. Der Übung halber soll aber trotzdem an dieser Stelle eine β-Fehler - basierte Kalkulation des **Stichprobenumfangs** vorgenommen werden. Die Firma SPSS unterstützt solche Berechnungen im Zusatzprogramm **SamplePower**, das uns leider

[1] Hier handelt es sich um einen statistischen Terminus, der nur zufällig mit unserer allgemeinpsychologischen Hypothese den Wortbestandteil *alternativ* gemeinsam hat.

nicht zur Verfügung steht. Stattdessen verwenden wir das exzellente Power-Analyse-Programm **G*Power 3.1** (Faul et al. 2007, 2009), das für MacOS und Windows kostenlos über folgende Webseite zu beziehen ist:

http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

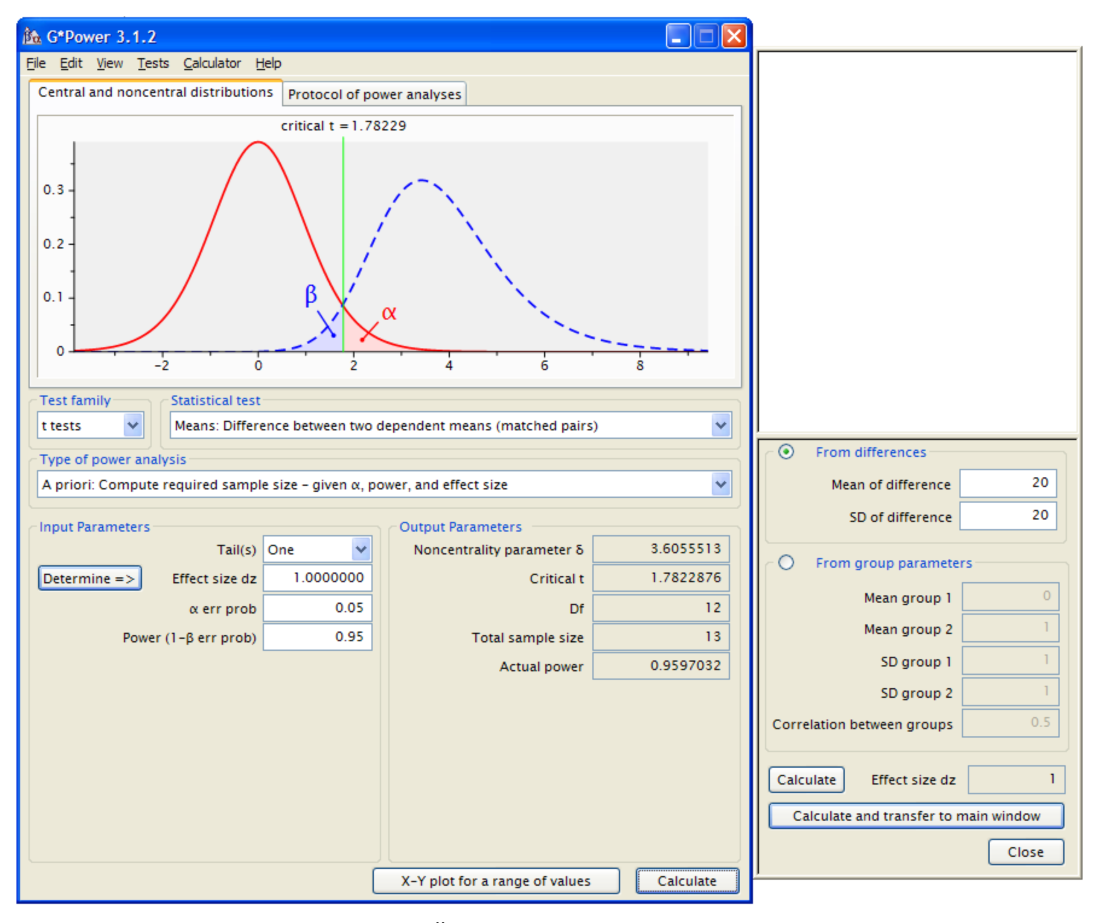
Auf den Pool-PCs der Universität Trier unter dem Betriebssystem Windows lässt sich G*Power 3.1 über folgende Programmgruppe starten

Start > Alle Programme > Wissenschaftliche Programme > GPower

Wir wählen

- **Test family**: **t-Tests**
- **Statistical test**: **Means: Difference between two dependent means**
- **Type of power analysis**: **A priori**

und öffnen über den Schalter **Determine** ein Zusatzfenster, um die Effektstärke in der Population aufgrund theoretischer Annahmen und/oder bisheriger empirischer Erfahrungen festlegen zu können:



Unsere KFA-Hypothese handelt vom *Ärgerzuwachs* aufgrund der kontrafaktischen Alternative und kann über die Differenz der beiden Ärgermessungen beurteilt werden. Wir verwenden in G*Power 3.1 diese Sichtweise, um die **Effektstärke** *dz* bequem festlegen zu können. Die Effektstärke ist (wie beim letztlich zugrunde liegenden Einstichproben - t-Test) folgendermaßen definiert (vgl. z.B. Wentura 2004, S. 4):

$$dz := \frac{\mu}{\sigma}$$

Darin sind:

- μ Mittelwert des betrachteten Merkmals (hier: Ärger*zuwachs*) in der Population
- σ Standardabweichung des Merkmals in der Population

Als mittleren Ärgerzuwachs (Hauptparameter der KFA-Hypothese) erwarten wir ca. 20°. Als Ärgerzuwachs-Standardabweichung (Nebenparameter der KFA-Hypothese) erwarten wir aufgrund bisheriger Studien ebenfalls einen Wert von ca. 20. Mit dem Schalter

Calculate and transfer to main window

befördern wir die resultierende Effektstärke von 1,0 in das G*Power-Hauptfenster.

Dort wählen wir …

- einen gerichteten Test (**Tail(s)**: **One**)
- einen akzeptierten α-Fehler (**α err prob**) der Größe 0,05
  Bei der Eingabe ist zu beachten, dass G*Power nur den *Punkt* als Dezimaltrennzeichen akzeptiert.
- eine gewünschte Teststärke (**Power**) von 0,95, also einen β-Fehler von 0,05

Nach einem Mausklick auf den Hauptfensterschalter

Calculate

erhalten wir das beruhigende Ergebnis, dass lediglich eine Stichprobe mit 13 Fällen erforderlich ist. Sofern ein Effekt von der angenommenen Stärke vorhanden ist, werden wir ihn mit großer Wahrscheinlichkeit entdecken, weil die Kursstudie in der Regel mehr als 30 Fälle enthält.

1.3.3 Eine differentialpsychologische Hypothese

Neben der zentralen KFA-Hypothese soll in unserer Studie die folgende, auf Überlegungen von Scheier & Carver (1985) zurückgehende Hypothese überprüft werden:

> **Der durch ein negatives Ereignis ausgelöste Ärger wird durch dispositionellen Optimismus gedämpft.**

Begründung: Dispositioneller Optimismus (im Sinne generalisierter positiver Ergebniserwartungen) führt zur Verwendung günstiger Bewältigungsstrategien (z.B. positive Reinterpretation von negativen Erfahrungen).

Während die allgemeinpsychologische KFA-Hypothese für eine beliebig aus der Allgemeinbevölkerung herausgegriffene Person einen Effekt postuliert, geht es hier um Differentialpsychologie, also um Verhaltensunterschiede in Folge von relativ stabilen Personmerkmalen.

Als Quasiereignis soll der schon zur Prüfung der allgemeinpsychologischen Hypothese verwendete imaginierte Schadensfall dienen (Fragebogenteil 2, siehe unten). Das arithmetische Mittel der für beide Situationsvarianten angegebenen Ärgerausprägungen soll als Ärgermaß dienen. Zur Erfassung von dispositionellem Optimismus wird der von Scheier & Carver (1985) entwickelte *Life Orientation Test* (LOT) eingesetzt (siehe Fragebogenteil 3). Wie aus den Antworten auf die zwölf Fragen dieses Tests ein Optimismus-Messwert zu ermitteln ist, wird später erläutert. Wir gehen jedenfalls davon aus, dass diese Messmethode annähernd Intervallniveau besitzt.

Nach dieser **Operationalisierung** der theoretischen Begriffe kann die folgende **empirisch prüfbare Alternativhypothese** formuliert werden:

> **Je höher der LOT-Wert einer Versuchsperson, desto weniger Ärger berichtet sie im Mittel für den imaginierten Schadensfall.**

Die Nullhypothese ergibt sich durch Negation der Alternativhypothese und muss daher nicht notiert werden.

Weil die Messungen zum Ärger und zum Optimismus (hoffentlich) Intervallskalenniveau besitzen, kann die differentialpsychologische Hypothese mit einer **einfachen (bivariaten) linearen Regressionsanalyse** geprüft werden, sofern deren Modell- und Verteilungsvoraussetzungen erfüllt sind. Bei der bivariaten linearen Regression kann man tatsächlich von einem Modell sprechen. Seine Modellgleichung und ein Streudiagramm mit Stichprobendaten zu einer analogen Fragestellung waren schon in Abschnitt 1.1 zu sehen.

Nun lässt sich die statistisch zu prüfende Alternativhypothese noch präziser formulieren:

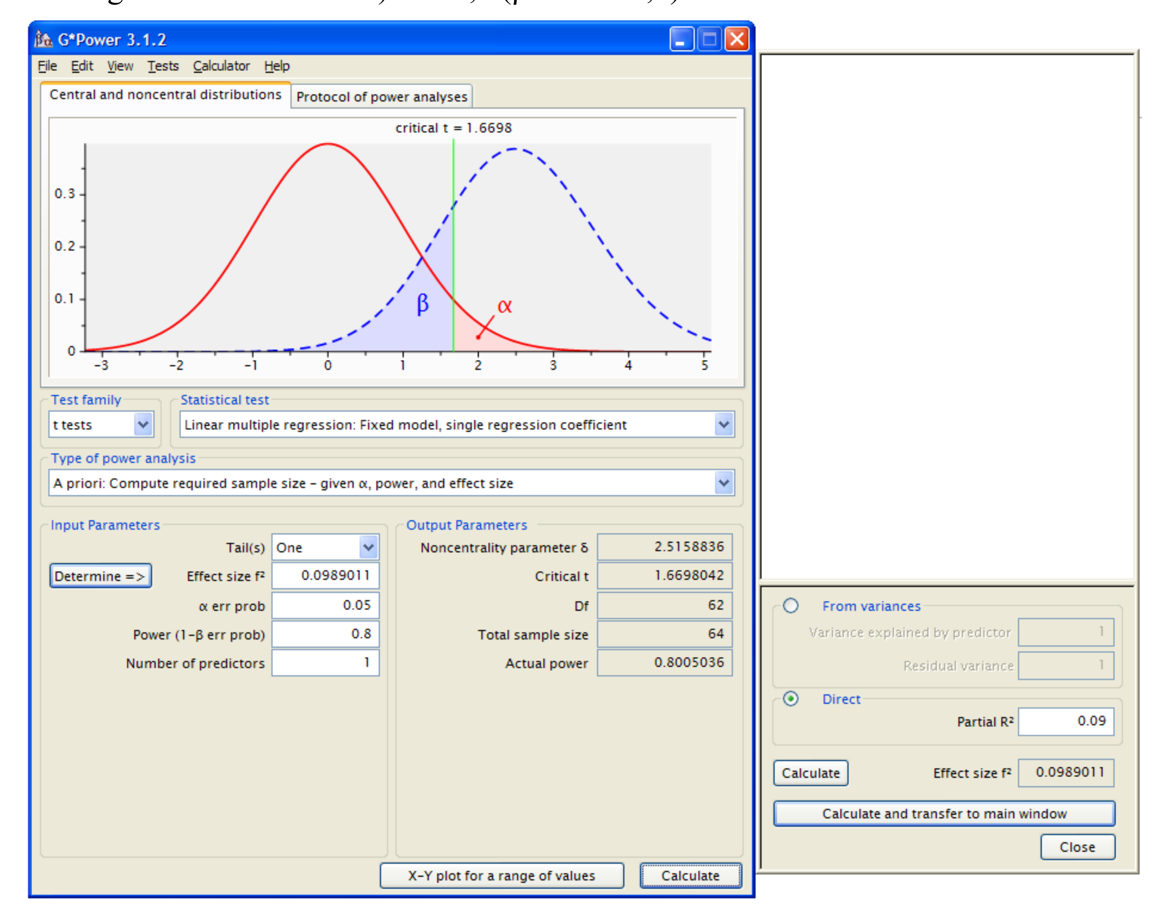
> **In der linearen Regression von Ärger auf Optimismus ergibt sich ein negativer Steigungskoeffizient (β_1).**

Die Hypothese ist *gerichtet (einseitig)* formuliert und soll mit einem α-Fehler – Risiko von 5% mit dem t-Test zum Regressionskoeffizienten β_1 geprüft werden.

Zur Berechnung der erforderlichen Stichprobengröße wählen wir im Teststärkenanalyseprogramm G*Power 3.1 (vgl. Abschnitt 1.3.2):

- **Test family**: **t-Tests**
- **Statistical test**: **Linear Multiple Regression: Fixed model, single regression coefficient**
- **Type of power analysis**: **A priori**

Für den geplanten einseitigen t-Test mit einem α-Fehler – Risiko von 5% in einem Modell mit einem Prädiktor wählen wird die von Cohen (1977, S. 56) als Standardwert empfohlene Power (Entdeckungswahrscheinlichkeit) von 0,8 (β-Fehler 0,2):



Das von G*Power verwendete Effektstärkemaß f^2 steht in folgender Beziehung zum Determinationskoeffizienten R^2 der linearen Regression (Anteil der erklärten Kriteriumsvarianz, Quadrat der multiplen Korrelation der Prädiktoren mit dem Kriterium):

$$f^2 = \frac{R^2}{1 - R^2}$$

Wir nehmen einen Determinationskoeffizienten von 0,09 (eine LOT-Ärger - Korrelation von -0,3) und damit eine Effektstärke von ca. 0,1 an:

$$\frac{0{,}09}{1 - 0{,}09} = \frac{0{,}09}{0{,}91} \approx 0{,}1$$

In G*Power öffnen wir mit dem Schalter **Determine** das Seitenfenster zur Effektstärkenspezifikation, wählen dort die Option **Direct**, tragen im Textfeld **Partial R^2** den Wert 0,09 ein, und übernehmen den resultierenden f^2-Wert mit dem Schalter **Calculate and transfer to main window**.

Aus unserer Problembeschreibung resultiert ein erforderlicher Stichprobenumfang von 64 Fällen. Weil die Kursstichprobe in der Regel kleiner ist, stehen unsere Chancen einen Effekt von der vermuteten Stärke zu entdecken also eher schlecht. Bei einer gewünschten Power von 0,95 (β-Fehler 0,05) werden sogar 111 Fälle benötigt. In einem realen Forschungsprojekt zur Klärung der differentialpsychologischen Hypothese müsste der Stichprobenumfang folglich erhöht werden.

Bei einem *zweiseitigen* Test werden bei der oben angenommenen Effektstärke und $\alpha = \beta = 0{,}05$ bereits 134 Fälle benötigt. Wer den Unterschied zwischen gerichteten und ungerichteten Hypothesen ignoriert und mit dem bei EDV-Programmen üblicherweise voreingestellten zweiseitigen Test arbeitet, muss also einen erhöhten Aufwand bei der Datenerhebung betreiben bzw. verliert bei identischem Stichprobenumfang in erheblichem Umfang an Teststärke.

Damit ein *starker* Effekt (R^2= 0,25 bzw. f^2 = 0,33) bei einseitiger Testung zum α-Niveau 0,05 mit einer Wahrscheinlichkeit von 0,8 zu einem signifikanten Ergebnis führt, sind nur 21 Fälle erforderlich, so dass auch in der relativ kleinen Kursstichprobe noch Anlass zur Hoffnung besteht.

1.3.4 Zum Einfluss demographischer Merkmale

Auf die Erfassung demographischer Merkmale (siehe Fragebogenteil 1) kann man in keiner Studie mit Personen als Beobachtungseinheiten verzichten, auch wenn sich keine expliziten Hypothesen darauf beziehen. Man benötigt sie auf jeden Fall zur Beschreibung der Stichprobe, damit sich später die Leser(innen) von Berichten ein Urteil über die Interpretier- bzw. Generalisierbarkeit der Ergebnisse bilden können. Wir werden darüber hinaus einige demographische Merkmale auf Zusammenhänge mit unseren zentralen Projektvariablen untersuchen. Insofern finden sich auch in unserem überwiegend konfirmatorisch (hypothesenprüfend) angelegten Projekt einige explorative Elemente.

1.3.5 Zu Übungszwecken erhobene Merkmale

Rein zu Übungszwecken und ohne inhaltlichen Bezug zu den Fragestellungen des Projekts sollen zusätzlich folgende Informationen erhoben werden:

- Größe und Gewicht (siehe Fragebogenteil 1)
  Mit diesen Merkmalen lassen sich manche statistische Verfahren gut demonstrieren. Außerdem sorgen sie für das Auftreten gebrochener Zahlen in unseren Daten.
- Motive zur Kursteilnahme (siehe Fragebogenteil 4)
  Hier geht es um die Strukturierung der Informationen, die aus Mehrfachwahlfragen und offenen Fragen resultieren.

1.3.6 Der Fragebogen

1) Angaben zur Person

Geschlecht	Frau ❐	Mann ❐
Geburtsjahr		
Fachbereich		
Körpergröße	__,__ __ m	
Körpergewicht	__ __ kg	

2) Fragen zur Reaktion in ärgerlichen Situationen

Versetzen Sie sich bitte möglichst gut in folgende Situation:

Herr Meier und Herr Schulze waren mit demselben Taxi auf dem Weg zum Flughafen. Sie sollten zur selben Zeit, aber mit verschiedenen Maschinen abfliegen. Durch einen Stau kommen sie erst eine halbe Stunde nach der planmäßigen Abflugzeit am Flughafen an.

***Herr Meier** erfährt, dass seine Maschine pünktlich vor einer halben Stunde gestartet ist.*

***Herr Schulze** erfährt, dass seine Maschine Verspätung hatte und erst vor zwei Minuten gestartet ist.*

Wie sehr würden Sie sich **ärgern**, wenn Sie in der Situation von ...

Herrn Meier wären?	0	10	20	30	40	50	60	70	80	90	100
Herrn Schulze wären?	0	10	20	30	40	50	60	70	80	90	100

Betrachten Sie bitte die Antwortskala als "Ärgerthermometer".

3) Aussagen zur Selbsteinschätzung

Teilen Sie bitte für die folgenden Selbstbeschreibungen durch Ankreuzen einer Antwortkategorie mit, inwiefern die Aussagen auf Sie persönlich zutreffen.

	völlig falsch	falsch	unent-schieden	stimmt	stimmt genau
1. Auch in unsicheren Zeiten rechne ich im Allgemeinen damit, dass sich alles zum Besten wendet.	- -	-	0	+	+ +
2. Ich kann mich leicht entspannen.	- -	-	0	+	+ +
3. Wenn etwas schief gehen kann, dann passiert es mir auch.	- -	-	0	+	+ +
4. Bei allem sehe ich stets die negative Seite.	- -	-	0	+	+ +
5. Ich blicke kaum einmal mit Zuversicht in die Zukunft.	- -	-	0	+	+ +
6. Ich bin gern mit Freunden zusammen.	- -	-	0	+	+ +
7. Ich muss mich immer mit etwas beschäftigen.	- -	-	0	+	+ +
8. Ich habe stets die Hoffnung, dass die Dinge in meinem Sinne gehen.	- -	-	0	+	+ +
9. Die Dinge laufen immer so, wie ich es mir wünsche.	- -	-	0	+	+ +
10. Ich bin nicht leicht aus der Ruhe zu bringen.	- -	-	0	+	+ +
11. Ich glaube an den sprichwörtlichen "Silberstreifen am Horizont".	- -	-	0	+	+ +
12. Dass mir einmal etwas Gutes widerfährt, damit rechne ich kaum.	- -	-	0	+	+ +

4) Ihre Motive für die Teilnahme am SPSS-Kurs

a) Kreuzen Sie bitte in der folgenden Liste möglicher Motive für die Teilnahme am SPSS-Kurs alle für Sie zutreffenden Aussagen an und/oder nennen Sie Ihre sonstigen Motive.

Ich möchte SPSS kennen lernen, ...

- ❐ um eine eigene empirische Studie damit auszuwerten.
- ❐ weil in vielen Stellenanzeigen SPSS-Kenntnisse verlangt werden.
- ❐ weil ich mich um eine Stelle als EDV-Hilfskraft in der Forschung bewerben will (HIWI-Job).
- ❐ weil ich mich für EDV interessiere und ein modernes Programm kennen lernen möchte.
- ❐ weil ich mich für Statistik interessiere und mit Auswertungsverfahren experimentieren möchte.

❐ Andere Motive: ______________________________

____

b) Möchten Sie im Kurs bestimmte statistische Methoden besonders gerne üben? Ja ❐ Nein ❐

Wenn „Ja“, welche? ______________________________

____

____

1.4 Strukturierung und Kodierung der Daten

Wir werden die mit unserem Fragebogen erhobenen Informationen später manuell mit dem SPSS-Dateneditor erfassen und erstellen daher einen **Kodierplan** mit genauen Handlungsanweisungen für die Erfassung. Dabei müssen wir uns auch mit den Voraussetzungen beschäftigen, die SPSS für die Aufnahme unserer Daten bereitstellt. Diese sind in erster Linie durch die Logik der empirischen Forschung und nur in geringem Ausmaß durch EDV-Restriktionen festgelegt.

Bei der automatischen Erhebung bzw. Erfassung (Online-Formular, Daten-Scanner) wird kein Kodierplan als Arbeitsvorschrift für Datenerfasser benötigt, jedoch kann auch hier eine Dokumentation der Daten nützlich sein (z.B. für die Kooperation in einer Arbeitsgruppe). Die in Abschnitt 1.4 behandelten Fragen werden bei den automatischen Methoden teilweise bei der Datendeklaration gegenüber der Umfrage- bzw. Scanner-Software geregelt und teilweise von der Software entschieden. Bei manchen Aufgaben sind Urteilsvermögen und Handarbeit eines Menschen durch keine Software zu ersetzen, z.B. bei der Behandlung der Antworten auf offene Fragen (siehe Abschnitt 1.4.2.4). Insgesamt kann der Abschnitt 1.4 auch solchen Lesern zur Lektüre empfohlen werden, die zu einer Online- oder Scanner-Lösung tendieren.

1.4.1 Fälle und Merkmale in SPSS

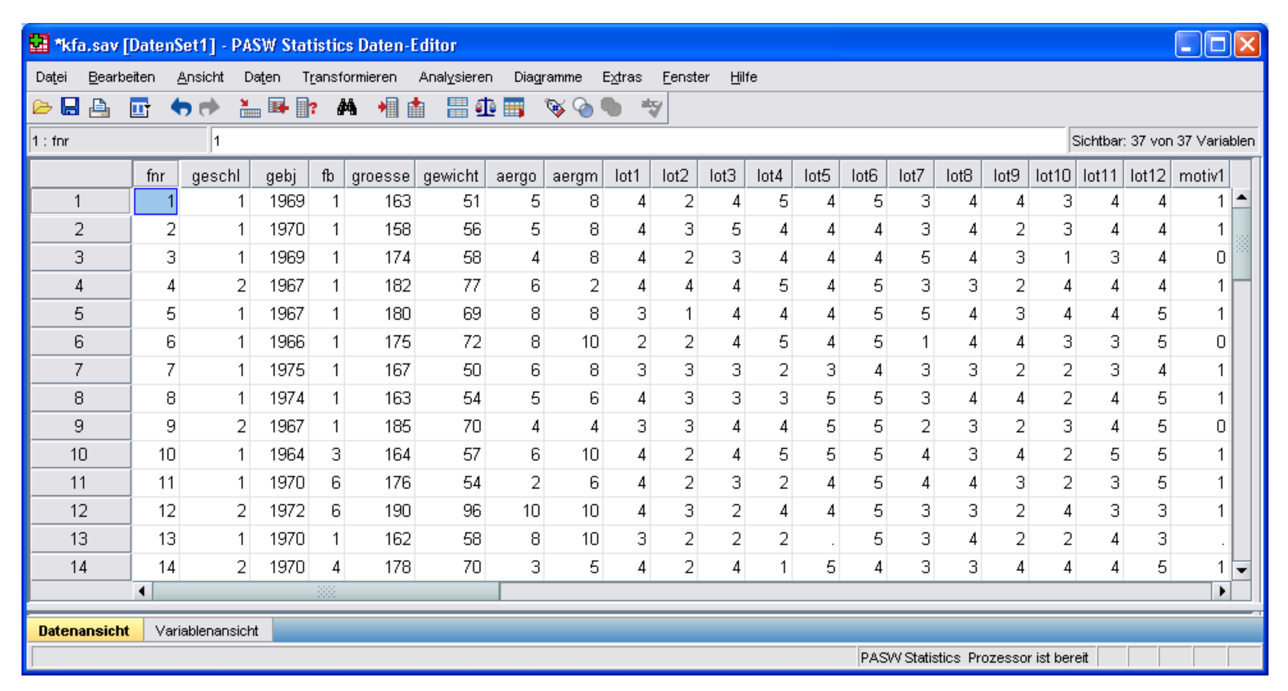
Wir haben oben bereits daran erinnert, dass in einer empirischen Studie bei den einbezogenen **Fällen** bzw. **Beobachtungseinheiten** die Ausprägungen etlicher **Merkmale** festgestellt werden. Nun wollen wir uns ansehen, wie die Merkmalsausprägungen der Fälle im SPSS-System gespeichert werden. Die ganz konkrete Demonstration von KFA-Beispieldaten im **SPSS-Dateneditorfenster** wird das Verständnis der anschließenden, wieder eher allgemein-methodologisch geprägten, Ausführungen sicher unterstützen. U.a. werden dabei auch einige zentrale Begriffe des SPSS-Systems erläutert:

a) Variable

Der Begriff *Variable* wird in der Literatur zur statistischen Datenanalyse häufig synonym zu *Merkmal* gebraucht. Wir wollen ihn SPSS-konform in einer etwas technischeren Bedeutung verwenden: Schreibt man für ein Merkmal die Ausprägungen aller Fälle in der Stichprobe untereinander, so entsteht ein Spaltenvektor, und einen solchen Spaltenvektor wollen wir als *Variable* bezeichnen. Zwar resultieren Variablen meist (wie gerade beschrieben) aus jeweils *einem* Merkmal, doch kann z.B. das Bemühen um eine rationelle Datenerfassung zu Ausnahmen führen. In Kürze wird eine Technik vorgeschlagen, die zur Erfassung von 100 Merkmalen mit Hilfe von fünf Variablen führt.

b) Datenmatrix und Dateneditor

Schreibt man alle Variablen nebeneinander, so entsteht die (Fälle × Variablen) - Datenmatrix (Datentabelle). Sie kann in einem Fenster des SPSS-Dateneditors aufgebaut und dort auch während der laufenden Auswertungsarbeit ständig eingesehen oder bearbeitet werden. Die folgende Abbildung zeigt ein Dateneditorfenster mit KFA-Beispieldaten aus einem früheren SPSS-Kurs:



	fnr	geschl	gebj	fb	groesse	gewicht	aergo	aergm	lot1	lot2	lot3	lot4	lot5	lot6	lot7	lot8	lot9	lot10	lot11	lot12	motiv1
1	1	1	1969	1	163	51	5	8	4	2	4	5	4	5	3	4	4	3	4	4	1
2	2	1	1970	1	158	56	5	8	4	3	5	4	4	4	3	4	2	3	4	4	1
3	3	1	1969	1	174	58	4	8	4	2	3	4	4	4	5	4	3	1	3	4	0
4	4	2	1967	1	182	77	6	2	4	4	4	5	4	5	3	3	2	4	4	4	1
5	5	1	1967	1	180	69	8	8	3	1	4	4	4	5	5	4	3	4	4	5	1
6	6	1	1966	1	175	72	8	10	2	2	4	5	4	5	1	4	4	3	3	5	0
7	7	1	1975	1	167	50	6	8	3	3	3	2	3	4	3	3	2	2	3	4	1
8	8	1	1974	1	163	54	5	6	4	3	3	3	5	5	3	4	4	2	4	5	1
9	9	2	1967	1	185	70	4	4	3	3	4	4	5	5	2	3	2	3	4	5	0
10	10	1	1964	3	164	57	6	10	4	2	4	5	5	5	4	3	4	2	5	5	1
11	11	1	1970	6	176	54	2	6	4	2	3	2	4	5	4	4	3	2	3	5	1
12	12	2	1972	6	190	96	10	10	4	3	2	4	4	5	3	3	2	4	3	3	1
13	13	1	1970	1	162	58	8	10	3	2	2	2	.	5	3	4	2	2	4	3	.
14	14	2	1970	4	178	70	3	5	4	2	4	1	5	4	3	3	4	4	4	5	1

Jede Variable, d.h. jede Spalte der Datenmatrix, besitzt einen eindeutigen **Variablennamen**, über den sie bei der Anforderung statistischer oder graphischer Analysen angesprochen werden kann.

Nachdem Sie einen exemplarischen Eindruck vom *Ziel* der Strukturierungs- und Kodierungsbemühungen gewonnen haben, werden wir nun einige Details behandeln und einen Kodierplan für unser Projekt erstellen. Dabei soll u.a. angestrebt werden, den Aufwand und die Fehlergefahr bei der Datenerfassung möglichst gering zu halten.

1.4.2 Strukturierung

Welche SPSS-Variablen im oben besprochenen Sinn sollen zur Aufnahme der mit unserem Fragebogen erfassten Informationen definiert werden? Obwohl die Antwort auf diese Frage trivial zu sein scheint, sind doch zu einigen Themen kurze Erläuterungen angebracht.

1.4.2.1 Variablen zur Fallidentifikation

Über die empirischen Variablen hinaus sollten in die Datenmatrix stets organisatorische Variablen aufgenommen werden, die eine Relation zwischen den schriftlichen oder sonstigen Untersuchungsdokumenten eines Falles und seinen Daten im Rechner herstellen. Eine solche Korrespondenz ist für eventuelle spätere Kontrollen oder Korrekturen der Daten unbedingt erforderlich. Meist verwendet man für diesen Zweck eine *einzelne* Variable, die z.B. FNR (für *Fallnummer*) genannt werden kann. Natürlich muss die Fallidentifikation auch auf den schriftlichen oder sonstigen Untersuchungsdokumenten eingetragen werden. Bei personbezogenen Daten wählt man aus Datenschutzgründen zur Fallidentifikation z.B. eine zufällig vergebene Nummer ohne jeden Bezug zu den Personalien.

Möglicherweise erscheint Ihnen das Eintippen einer Identifikationsvariablen sinnlos, weil im SPSS-Dateneditor (siehe Abbildung in Abschnitt 1.4.1) die Zeilen bzw. Fälle ohnehin fortlaufend nummeriert sind. Die Nummern der Datenfensterzeilen stellen jedoch die gewünschte Korrespondenz zwischen den Datensätzen im Rechner und den nummerierten schriftlichen Untersuchungsunterlagen *nicht zuverlässig* her. Die Nummerierung der Datenfensterzeilen kann sich nämlich leicht ändern, z.B. wenn ein Sortieren der Fälle nötig wird, oder wenn Fälle gelöscht oder eingefügt werden.

1.4.2.2 *Abgeleitete Variablen gehören nicht in den Kodierplan*

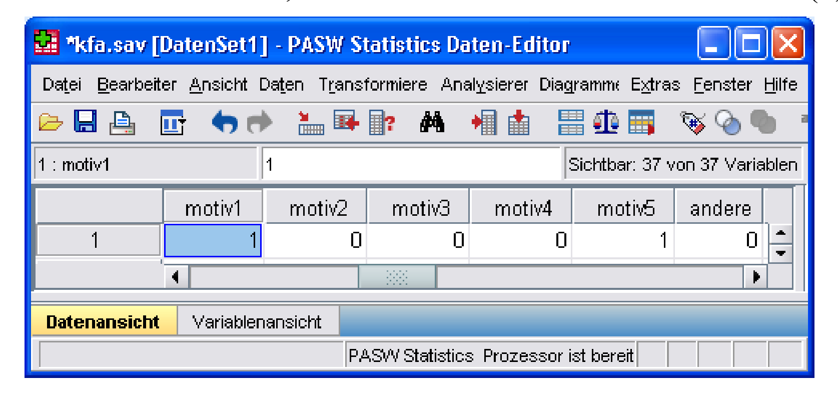
Häufig sind in einem Forschungsprojekt nicht nur die direkt erfassten *Rohvariablen* von Interesse, sondern auch darauf aufbauende Variablen. Im KFA-Projekt soll etwa der Optimismus der Untersuchungsteilnehmer durch ihre mittlere Antwort auf die LOT-Fragen geschätzt werden. SPSS verfügt über leistungsfähige Dialogboxen bzw. Befehle zur Berechnung neuer Variablen aus bereits vorhandenen, so dass derartige Routinearbeiten keinesfalls bei der Datenerfassung (z.B. per Taschenrechner) erledigt werden sollten. Freilich müssen nach diesem Vorschlag *alle* Ausgangsvariablen aufgenommen werden, was aber vielfach ohnehin erforderlich ist (z.B. zur Überprüfung messtechnischer Eigenschaften der Ausgangsvariablen). Erfassen Sie also ausschließlich Rohvariablen, und führen Sie alle erforderlichen Transformationen später mit SPSS-Techniken durch. Wir werden uns im weiteren Kursverlauf mit den SPSS-Transformationsmethoden ausführlich beschäftigen.

1.4.2.3 *Mehrfachwahlfragen*

Im Teil 4a unseres Fragebogens teilen die Untersuchungsteilnehmer für fünf konkrete Motive und eine Restkategorie mit, ob sie bei ihrer Entscheidung für die Kursteilnahme relevant waren. Damit erfahren wir von jeder Person sechs eigenständige Merkmalsausprägungen und benötigen (ohne Komprimierungsverfahren, siehe unten) folglich in der SPSS-Datentabelle sechs Variablen, um die Antworten aufzunehmen, die wir z.B. durch die Zahlen Eins (für *trifft zu*) und Null (für *trifft nicht zu*) kodieren können. Beim Umgang mit einer solchen Mehrfachwahlfrage müssen Sie sich vor allem vor dem aussichtslosen Versuch hüten, die Informationen zu allen Merkmalen in *eine* Variable zu verpacken. Dies käme dem unsinnigen Versuch gleich, *mehrere* Werte (z.B. Zahlen) in *eine* Zelle der SPSS-Datenmatrix einzutragen.

1.4.2.3.1 Vollständige Sets aus dichotomen Variablen

In unserem Beispiel führt also eine Mehrfachwahlfrage zu sechs dichotomen SPSS-Variablen, die jeweils die Information darüber enthalten, ob ein bestimmtes Motiv (bzw. ein sonstiges Motiv) vorlag oder nicht. Das folgende Datenfenster zeigt die sechs Variablen, hier mit den Namen MOTIV1 bis MOTIV5 und ANDERE, bei einem Fall mit dem Antwortmuster (1,0,0,0,1,0):



Wir werden in Abschnitt 13 ein so genanntes **Mehrfachantworten-Set** bestehend aus diesen sechs Variablen definieren und mit seiner Hilfe eine gemeinsame Auswertung der Variablen vornehmen. An dieser Stelle müssen Sie jedoch unbedingt akzeptieren, dass wir es mit *sechs* Merkmalen bzw. Variablen zu tun haben, die eine gewisse Verwandtschaft und ein gemeinsames dichotomes Format besitzen.

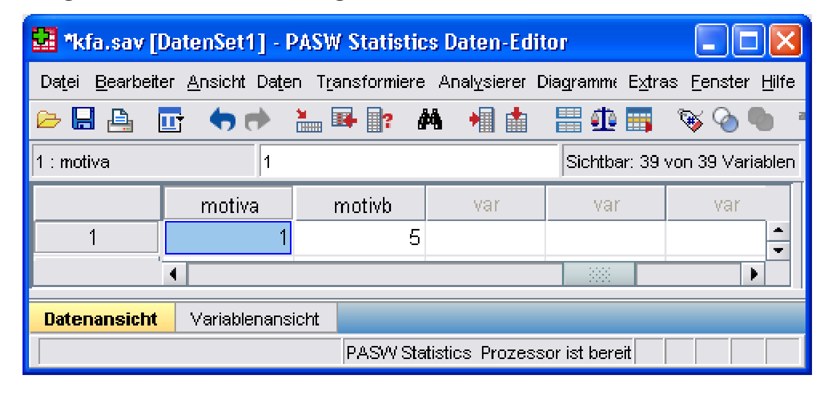
1.4.2.3.2 Sparsame Sets aus kategorialen Variablen

Das im letzten Abschnitt beschriebene Standardverfahren zur Übersetzung einer Mehrfachwahlfrage in SPSS-Variablen ist angemessen, sofern nicht zu viele Antwortmöglichkeiten im Spiel sind. Wenn Sie etwa eine Liste mit 100 möglichen Freizeitaktivitäten präsentieren, dann führt das Schema zur Definition von 100 SPSS-Variablen. Unter der Annahme, dass jeder einzelne Untersuchungsteilnehmer maximal sieben verschiedene Optionen wählen wird, ist das Schema bei der Datenerfassung recht unpraktisch. Für solche Situationen bietet sich ein alternatives Vorgehen an, das im eben konstruierten Freizeitbeispiel lediglich sieben Variablen bzw. Spalten in der SPSS-Datentabelle benötigt.

Auch dieses Komprimierungsverfahren soll an unserem Motivbeispiel demonstriert werden, obwohl es hier (bei nur sechs Antwortmöglichkeiten) ungeeignet ist. Unter der Annahme, dass pro Person maximal *zwei* verschiedene Motive zutreffen werden, definiert man die beiden SPSS-Variablen MOTIVA und MOTIVB, die jeweils folgende Werte annehmen können:

1 für das Motiv *Eigene empirische Studie*
2 für das Motiv *Orientierung am Arbeitsmarkt*
3 für das Motiv *Bewerbung als EDV-Hilfskraft*
4 für das Motiv *Interesse an der EDV*
5 für das Motiv *Interesse an Statistik*
6 für andere Motive

Mit den Variablen MOTIVA und MOTIVB stehen für jede Person *zwei* Möglichkeiten zur Verfügung, um die Nummern von angekreuzten Motiven zu erfassen. Das Antwortmuster (1,0,0,0,1,0) wird folgendermaßen übertragen:



Im Prinzip kann man im Beispiel die beiden Werte Eins und Fünf auch in umgekehrter Reihenfolge eintragen (MOTIVA = 5, MOTIVB = 1). Wesentlich ist nur, dass die Nummer jedes angekreuzten Motivs bei einer Variablen als Wert auftritt. Von einer Person, die zwei Motive angekreuzt hat, wissen wir *nicht*, welchem Motiv sie die größte Bedeutung beimisst. Daher können auch die resultierenden Variablen eine solche subjektive Ranginformation nicht enthalten. Allerdings wird man beim Erfassen der Systematik halber so vorgehen, dass in MOTIVA die Nummer des ersten angekreuzten Motivs landet usw. (bei Anordnung von oben nach unten).

Wir sparen vier Variablen ein, wobei kein Informationsverlust eintritt, wenn tatsächlich pro Person maximal zwei Motive angekreuzt werden. Erweist sich ein sparsames Set während der Erfassung als unterdimensioniert, kann es bei Verwendung des SPSS-Dateneditors problemlos erweitert werden (z.B. um die Variable MOTIVC).

Auch bei der sparsamen Informationsanordnung kann man mit SPSS z.B. für jedes Motiv ermitteln, wie viel Prozent der Kursteilnehmer es angekreuzt haben. Vor einer solchen Auswertung ist wiederum ein Mehrfachantworten-Set zu definieren, diesmal bestehend aus den beiden Variablen MOTIVA und MOTIVB. Bei manchen Auswertungen ist es aber doch erforderlich, über Transformationsanweisungen aus dem sparsamen kategorialen Set das vollständige dichotome Set (mit *einer* Variablen pro Merkmal) herzustellen (siehe Abschnitt 13.4).

1.4.2.4 *Offene Fragen*

Offene Fragen lösen vielfältige und oft schwer strukturierbare Antworten aus, und es bleibt dann offen, ob und wie die Antworten in SPSS-Variablen übersetzt werden sollen. Ein Weg zur Systematisierung und Erfassung der Antworten besteht darin, eine **Kategorienliste** zu entwickeln, die möglichst kurz ist und trotzdem die Antworten aller Teilnehmer gut repräsentiert. Anschließend kann man bei jedem Fall die vorhandenen bzw. fehlenden Nennungen der Listenelemente analog zu den Antworten auf eine Mehrfachwahlfrage erfassen.

Im Fall unseres Fragebogenteils 4b ist also durch Inspektion der ausgefüllten Fragebögen aller Teilnehmer eine Liste mit speziell gewünschten statistischen Auswertungsverfahren erstellen, z.B. mit dem Ergebnis:

Regressionsanalyse
Kreuztabellenanalyse
Faktorenanalyse
Diskriminanzanalyse

Bei der Umsetzung in SPSS-Variablen wird man bei einer relativ kurzen Kategorienliste ein vollständiges Set mit dichotomen Variablen verwenden, ansonsten ein sparsames Set aus kategorialen Variablen (siehe Abschnitt 1.4.2.3). Aus der obigen vierelementigen Liste mit speziellen methodischen Interessen entsteht also ein vollständiges Set mit dichotomen Variablen, z.B. mit den Namen:

REG	für die Regressionsanalyse
KT	für die Kreuztabellenanalyse
FAKT	für die Faktorenanalyse
DA	für die Diskriminanzanalyse

Bei der Variablen REG ist eine Eins einzutragen, wenn ein Fall auf die offene Frage hin die Regressionsanalyse angegeben und damit sein Interesse an dieser Methode signalisiert hat. Anderenfalls wird eine Null notiert, die aber *nicht* als explizit bekundetes Desinteresse an der Regressionsanalyse zu interpretieren ist.

Beim Erstellen eines Kategoriensystems sind zu enge Kategorien (mit sehr geringer Häufigkeit) ebenso ungeeignet wie zu breite Kategorien (mit geringem Informationsgehalt). Vielfach wird man aber mit einer Restkategorie arbeiten (z.B. *sonstige Methoden*), um bei vertretbarem Aufwand möglichst alle Äußerungen berücksichtigen zu können.

Das beschriebene Vorgehen erfordert zum Erstellen der Kategorienliste eine (speziell bei großen Stichproben) lästige Vorauswertung der Fragebögen, die sich mit folgendem Trick vermeiden lässt: Man verwendet eine **dynamisch wachsende Kategorienliste** in Verbindung mit einem sparsamen Set kategorialer Variablen. In unserem Beispiel kann man sich über ein sparsames Set mit den drei Variablen METH1 bis METH3 darauf vorbereiten, für jeden Fall maximal drei spezielle Auswertungsinteressen festzuhalten. Die Kategorienliste wird erst während der Datenerfassung entwickelt, indem man bei jedem Fall entscheidet, welche bereits definierten oder neu in die Liste aufzunehmenden Kategorien er im Fragebogenteil 4b angegeben hat. Die Liste kann dynamisch um beliebig viele Kategorien erweitert werden, weil die drei Variablen beliebig viele verschiedene Werte als Kategoriennummern aufnehmen können. Selbstverständlich müssen die neu aufgenommenen Kategorien mit den vergebenen Nummern sorgfältig dokumentiert werden. Falls mehrere Personen an der Erfassung beteiligt sind, muss die eindeutige Zuordnung durch entsprechende Verabredungen sichergestellt werden.

Offene Fragen sind sicher vielfach sinnvoll, weil sie Informationen zutage fördern können, an die bei der Untersuchungsplanung niemand gedacht hat. Gelegentlich sind die Antworten jedoch so spärlich oder so schlecht strukturierbar, dass eine *statistische* Analyse nicht lohnend erscheint. So werden erfahrungsgemäß im Teil 4a des Beispielfragebogens kaum individuelle Motive zur Kursteilnahme angegeben, und wir ignorieren diese offene Frage im weiteren Projektverlauf.

1.4.3 Kodierung

Für jedes erhobene Merkmal muss festgelegt werden, wie die einzelnen Merkmalsausprägungen kodiert werden sollen. Dabei ist eine Kodierung durch einfach aufgebaute Werte anzustreben (z.B. durch positive, ganze Zahlen). Bei konkreten Überlegungen zur Kodierung müssen wir berücksichtigen, welche Variablentypen von SPSS unterstützt werden.

1.4.3.1 Die wichtigsten Variablentypen in SPSS

An dieser Stelle beschränken wir uns auf die wichtigsten Variablentypen, mit denen die meisten Projekte auskommen:

- **Numerische Variablen**
  Werte: reelle Zahlen
  Z.B. geeignet für die Merkmale:
  - Alter
  - Größe
  - Gewicht
- **Zeichenkettenvariablen (synonym: alphanumerische Variablen, String-Variablen)**
  Werte: Folgen von Zeichen (Buchstaben, Ziffern, Sonderzeichen), bis zur SPSS-Version 12 beschränkt auf eine maximale Länge von 255 Zeichen
  Z.B. geeignet für die Merkmale:
  - Familienname
  - Man könnte das Merkmal Geschlecht alphanumerisch kodieren mit den Werten **f** und **m**.
- **Datumsvariablen**
  Werte: Datumsangaben
  Z.B. geeignet für das Merkmal: Geburtsdatum

Anwendungsfälle für Datumsvariablen dürften in der Regel klar erkennbar sein. Ansonsten müssen Sie sich nur zwischen der numerischen und der alphanumerischen Kodierung entscheiden. Bei Merkmalen mit **mindestens ordinalem Skalenniveau** ist offensichtlich nur die numerische Kodierung sinnvoll. Bei Merkmalen mit **Nominalskalenniveau** hat man hingegen die Wahl zwischen numerischer und alphanumerischer Kodierung der Merkmalsausprägungen.

Beispiel Geschlecht:
- numerische Kodierung: **1** für Frauen, **2** für Männer
- alphanumerische Kodierung: **f** für Frauen, **m** für Männer

Beim Arbeiten mit SPSS empfiehlt es sich, auch nominalskalierte Merkmale numerisch zu kodieren, weil manche Auswertungsverfahren auch dort numerische Variablen verlangen, wo aus statistischer Sicht lediglich nominales Messniveau erforderlich ist (z.B. die Diskriminanzanalyse).

1.4.3.2 *Das Problem fehlender Werte*

Trotz aller Sorgfalt sind in fast jedem Forschungsprojekt bei manchen Fällen einige Variablenausprägungen unbekannt, z.B. wegen technischer Probleme oder wegen nachlässig ausgefüllter Fragebögen. Bei der Kodierungsplanung muss daher für alle betroffenen Variablen festgelegt werden, was an Stelle fehlender oder ungültiger Werte in die zugehörigen Zellen der Datenmatrix eingetragen werden soll. Diese Ersatzwerte bezeichnet man häufig als *MD-Indikatoren*, wobei *MD* für *Missing Data* steht. Wir beschränken uns bei der anschließenden Behandlung des Themas auf numerische Variablen.

1.4.3.2.1 Benutzerdefinierte MD-Indikatoren

Grundsätzlich kann jede Zahl als MD-Indikator für eine numerische Variable vereinbart werden, z.B. der Wert Neun bei einer verweigerten Auskunft über das in fünf Stufen erfasste Einkommen. Gelegentlich sind bei einer Variablen sogar mehrere MD-Indikatoren nötig, wobei z.B. ein erster Indikator signalisiert *Frage trifft nicht zu* und ein zweiter bedeutet *Keine auswertbare Antwort geliefert*.

Beispiel: Wenn die Besucher einer touristischen Einrichtung (z.B. Radwanderweg) per Rating-Skala nach der Zufriedenheit mit ihrer Ferienunterkunft befragt werden sollen, kann man die SPSS-Variable ZUFU definieren und dabei folgende Kodierungsregeln vereinbaren:

- Vorhandene Antworten werden durch die Zahlen 1 bis 5 kodiert.
- Ersatzwerte für fehlende Antworten:
    - Tagesgäste erhalten den MD-Indikator 8 (*Frage trifft nicht zu*).
    - Bei Übernachtungsgästen wird der MD-Indikator 9 vergeben.

Beachten Sie bei der Verwendung von benutzerdefinierten MD-Indikatoren folgende Regeln:

- Es ist klar, dass alle MD-Indikatoren einer Variablen außerhalb des validen Wertebereichs liegen müssen. So wäre z.B. die 99 kein geeigneter MD-Indikator für unsere Variable Körpergewicht (gemessen in kg).
- Wählen Sie möglichst prägnante oder extreme Werte (also z.B. bei einer Variablen mit den validen Werten 1 und 2 den MD-Indikator 9). Dies bewirkt warnend auffällige Ergebnisse, falls Fälle mit fehlenden Werten nicht ordnungsgemäß von einer Analyse ausgeschlossen wurden.
- Der Einfachheit halber sollte für alle Variablen mit ähnlichem Wertebereich derselbe MD-Indikator verwendet werden.

Wichtig: Für jede betroffene Variable müssen dem SPSS-System alle benutzerdefinierten MD-Indikatoren bekannt gemacht werden (siehe Abschnitt 3.2.2).

1.4.3.2.2 System-Missing (SYSMIS)

Neben den vom Benutzer variablenspezifisch vereinbarten MD-Indikatoren verwendet SPSS für *alle numerischen* Variablen automatisch einen weiteren MD-Indikator, der mit *System-Missing*, *systemdefiniert fehlend* oder *SYSMIS* bezeichnet wird. Er kommt immer dann zum Einsatz, wenn SPSS auf eines der folgenden Probleme trifft:

- Im Dateneditor bzw. beim Lesen einer bereits vorhandenen Datendatei (z.B. im Textformat) findet SPSS im Feld einer als numerisch definierten Variablen unzulässige Zeichen oder überhaupt keinen Eintrag.
- Beim Neuberechnen einer Variablen per Transformationsanweisung (siehe unten) fehlt ein Argument, oder der Funktionswert ist nicht definiert (z.B. bei Division durch Null).

Gerade war u.a. zu erfahren, dass man beim Erfassen eines neuen Falles per SPSS-Dateneditor für eine Variable den Ersatzwert SYSMIS ganz einfach dadurch vereinbaren kann, dass man in die betroffene Zelle *nichts* einträgt.

Tipp: Bei der Datenerfassung mit dem SPSS-Dateneditor können Sie für numerische Variablen routinemäßig SYSMIS als MD-Indikator verwenden, bei Bedarf ergänzt durch zusätzliche benutzerdefinierte MD-Indikatoren. Man kann SYSMIS bequem dadurch vereinbaren, dass man die betroffene Zelle unverändert lässt. Weil SPSS den Ersatzwert SYSMIS automatisch richtig versteht, ist eine Deklaration nicht nötig und kann daher auch nicht vergessen werden.

Im Datenfenster und in der Ergebnisausgabe wird SYSMIS durch einen Punkt dargestellt (siehe Abbildung in Abschnitt 1.4.1, Variable LOT5 bei Fall 13).

1.4.3.2.3 Fehlende Werte bei Mehrfachwahl-Fragen und offenen Fragen

Nachdem der Sinn und die Verwendung von MD-Indikatoren geklärt sind, geht es in diesem Abschnitt um eine spezielle Interpretationsunsicherheit im Zusammenhang mit fehlenden Werten, die bei Mehrfachwahlfragen aus der Verwendung eines Teilnehmer-freundlichen Antwortformats resultieren kann. Im Fragebogenteil 4a zu den Motiven für die Kursteilnahme sorgt die sechste Ankreuzalternative (*Andere Motive*) durch Komplettieren der Antwortmöglichkeiten dafür, dass eine redliche Auskunftsperson mindestens eines der sechs Kästchen ankreuzen muss. Ohne diese Restkategorie könnten wir bei einem Fragebogen mit fünf leeren Motivkästchen folgende Möglichkeiten nicht unterscheiden:

- Bei der Person trifft tatsächlich keines der fünf vorgegebenen Motive zu.
- Die Person hat den Fragebogenteil 4a nicht bearbeitet (fehlende Daten).

Ursache für die Interpretationsunsicherheit ist offenbar das vereinfachte Antwortformat, das pro Motiv nur *ein* Kästchen vorsieht, statt jeweils ein Ja- *und* ein Nein-Kästchen vorzugeben. Damit ersparen wir den Untersuchungsteilnehmern zahlreiche Nein-Markierungen. Dies ist sinnvoll, damit deren Kooperationsbereitschaft nicht überstrapaziert wird, und die Fehlerquote gering bleibt.

Bei der offenen Frage in Teil 4b wird durch die vorgeschaltete Frage, ob überhaupt spezielle Methoden gewünscht sind, dafür gesorgt, dass bei Fragebögen ohne eingetragene Methodeninteressen folgende Möglichkeiten unterschieden werden können:

- Die Person hat kein Interesse an speziellen Auswertungsmethoden.
- Die Person hat den Fragebogenteil 4b nicht bearbeitet (fehlende Daten).

Durch das Bemühen um die Unterscheidbarkeit von verneinenden und fehlenden Antworten sollte das Fragebogendesign allerdings nicht zu umständlich bzw. pedantisch geraten.

1.4.3.2.4 Vereinfachung der Erfassung durch Datentransformationstechniken

Im Zusammenhang mit dem MD-Problem bei den Variablen zu unserem Fragebogenteil 4 wage ich nun einige Vorschläge, die zwar dem Datenerfasser das Leben erleichtern, aber zugegebenermaßen die Kursteilnehmer(innen) beim ersten Entwurf eines Kodierplans durch zusätzliche Überlegungen belasten. Bei der Mehrfachwahlfrage nach den Kursmotiven haben wir geschickt durch die sechste Ankreuzalternative *Andere Motive* dafür gesorgt, dass Personen mit fehlenden Werten sicher zu identifizieren sind. Wir könnten den Erfasser im Kodierplan beauftragen:

- Schreibe bei den Variablen MOTIV1 bis MOTIV5 und ANDERE den Wert Eins, wenn das zugehörige Kästchen markiert ist, sonst eine Null.
- Ist aber *keines* der sechs Kästchen markiert, dann versorge die Variablen MOTIV1 bis MOTIV5 und ANDERE mit dem MD-Indikator SYSMIS.

Die im zweiten Satz enthaltene Regel lässt sich mit (später anzuwendenden) SPSS-Transformationskommandos bequem automatisieren, so dass wir den Erfasser damit nicht belasten wollen. Damit wird die Lösung des MD-Problems zugunsten einer möglichst einfachen Erfassung in die spätere Projektphase der Datentransformation verschoben. Schlussendlich soll für die Variablen MOTIV1 bis MOTIV5 und ANDERE folgende Kodierung sichergestellt sein:

0	=	nein
1	=	ja
System-Missing	=	Wert unbekannt

Zur Erfassung der Informationen im Fragebogenteil 4b wollen wir eine dynamische Kategorienliste mit einem zugehörigem sparsamen Set kategorialer Variablen METH1 bis METH3 (vgl. Abschnitt 1.4.2.4) entwickeln. Der damit schon reichlich belastete Erfasser soll folgendermaßen vorgehen (bei Verwendung des SPSS-Dateneditors):

- Die Antwort auf die Frage, ob spezielle Methodenwünsche bestehen, wird konventionell in der Variablen SMG mit folgender Kodierungsvorschrift erfasst:

0	=	nein
1	=	ja
System-Missing	=	keine Antwort

- In die Dateneditorzellen zu den Variablen METH1 bis METH3 sollen die Kategoriennummern der gewünschten Methoden eingetragen werden. Bei weniger als drei Nennungen soll in den nicht benötigten Zellen nichts eingetragen werden, was zum MD-Indikator SYSMIS führt. Diese Regel erleichtert die Erfassung und hat noch einen weiteren Vorteil: Sollte sich herausstellen, dass zusätzliche Variablen METH4 etc. benötigt werden, können wir diese ergänzen, ohne bei bereits erfassten Fällen irgendwelche Ersatzwerte (z.B. Nullen) nachtragen zu müssen.

Bei den Variablen METH1 bis METH3 soll später mit SPSS-Transformationsanweisungen dafür gesorgt, dass ihre Ausprägungen zuverlässig folgendermaßen interpretiert werden können:

0	=	Von der *i*-ten ($i = 1,..,3$) Option zur Nennung einer interessierenden Methode wurde kein Gebrauch gemacht.
natürliche Zahl ≥ 1	=	Die Methode mit dieser Kategoriennummer wurde angegeben.
System-Missing	=	Wert unbekannt

Dazu müssen unter den verschiedenen Wertekonstellationen der Variablen SMG und METH1 bis METH3 folgende Anpassungen vorgenommen werden:

		Mindestens eine speziell interessierende Methode angegeben?	
		Ja	Nein
SMG	1	METH1 ... METH3: SYSMIS → 0 Bem.: Korrektes Antwortverhalten. Variablen zu nicht benutzten Optionen (gem. Kodierplan bisher auf SYSMIS) werden auf 0 gesetzt.	SMG: 1 → SYSMIS Bem.: Irreguläres Antwortverhalten. METH1 bis METH3 behalten SYMIS. SMG wird ebenfalls auf SYMIS gesetzt.
	0	SMG: 0 → 1 METH1 ... METH3: SYSMIS → 0 Bem.: Leicht irreguläres Antwortverhalten. Wir sind großzügig und setzen SMG auf 1.	METH1 ... METH3: SYSMIS → 0 Bem.: Korrektes Antwortverhalten. Die Variablen zu allen Optionen (gem. Kodierplan bisher auf SYSMIS) werden auf 0 gesetzt.
	SYSMIS	SMG: SYSMIS → 1 METH1 ... METH3: SYSMIS → 0 Bem.: Leicht irreguläres Antwortverhalten. Wir sind großzügig und setzen SMG auf 1 sowie die Variablen zu nicht benutzten Optionen auf 0.	Bem.: Irreguläres Antwortverhalten. Alle Variablen behalten den Wert SYSMIS.

Vermutlich kam beim Lesen der letzten Ausführungen wenig Freude auf. Das MD-Problem verursacht oft erheblichen Aufwand, wobei auch Ermessenentscheidungen gefragt sind. Jedenfalls sind die vorgeschlagenen Methoden zur Erfassung der Informationen aus dem Fragebogenteil 4 recht simpel und praktikabel.

1.4.3.3 Fehlerquellen bei der manuellen Datenerfassung minimieren

Wenn die Daten manuell erfasst werden, ist bei den Kodierungsvereinbarungen darauf zu achten, dass dem Erfasser keine zeitaufwändigen und fehleranfälligen Arbeiten zugemutet werden, z.B.:

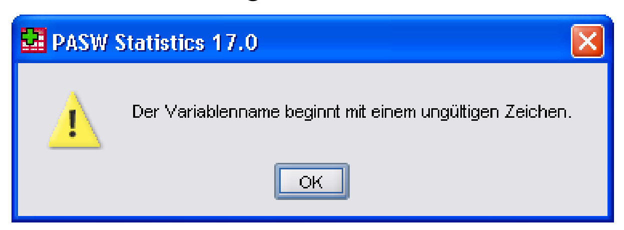
- Treten gebrochene Zahlen als Werte auf (z.B. bei unserer Frage nach der Körpergröße), so kann man durch Wechsel der Maßeinheit das lästige Dezimaltrennzeichen eliminieren.
  Beispiel: 1,65 m → 165 cm
- Bei bipolaren Skalen (z.B. bei unseren LOT-Fragen) empfiehlt sich eine Kodierung durch ausschließlich positive Werte z.B.:
  -- → 1
  - → 2
  0 → 3
  + → 4
  ++ → 5
  Durch die Vermeidung negativer Werte spart man Tipparbeit und macht keine Fehler durch vergessene Vorzeichen.
- Wurden einige Fragen aus messtechnischen Gründen umgepolt (negativ formuliert), was im KFA-Projekt bei einigen LOT-Fragen geschehen ist, so sollte diese Umpolung keinesfalls während der Erfassung rückgängig gemacht werden. Dies gelingt sehr viel bequemer und ohne Fehlerrisiko mit den Transformationsmöglichkeiten von SPSS (siehe unten).

1.4.3.4 SPSS-Variablennamen

Es empfiehlt sich, an dieser Stelle auch schon SPSS-Namen für die Variablen festzulegen und in den Kodierplan (siehe Abschnitt 1.4.3.5) aufzunehmen. Dabei sind die SPSS-Regeln für Variablennamen zu beachten:

- Maximal 64 Zeichen
  Die frühere Beschränkung von SPSS-Variablennamen auf acht Zeichen ist seit der Version 12 überwunden, doch sollte man sich weiterhin möglichst kurz fassen. Lange Namen belegen viel Platz (z.B. in der Kopfzeile des Dateneditorfensters) und sind beim Einsatz von SPSS-Syntax (siehe unten) recht umständlich.
- Das erste Zeichen muss ein Buchstabe sein.
- An den restlichen Positionen sind folgende Zeichen zugelassen: Buchstaben, Ziffern sowie die Symbole @, #, _ und $. Von der zweiten bis zur vorletzten Position ist außerdem der Punkt erlaubt.
- Aus den eben genannten Regeln ergibt sich insbesondere, dass Leerzeichen in Variablennamen verboten sind.
- Die von älteren SPSS-Versionen verschmähten Umlaute in Variablennamen werden mittlerweile akzeptiert. Seit der SPSS-Version 16 sind auch beim Übergang zu einem alternativen Betriebssystem keine Zeichensatzprobleme bei Variablennamen mehr zu befürchten. Trotzdem werden in diesem Manuskript mit Rücksicht auf ältere SPSS-Versionen Umlaute in Variablennamen vermieden.
- Die folgenden Schlüsselwörter der SPSS-Kommandosprache (siehe unten) dürfen nicht als Variablennamen verwendet werden: ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO, WITH.
- Die Groß-/Kleinschreibung ist irrelevant hinsichtlich der *Identifikation* von Variablen, jedoch verwendet SPSS bei *Ausgaben* die Schreibweise aus der Variablendeklaration. Ist zu einer Variablen allerdings ein Variablenlabel (siehe unten) definiert, erscheint dieses in der Ausgabe an Stelle des Namens. Wir schreiben in SPSS die Variablennamen aus Bequemlichkeitsgründen in Kleinbuchstaben. In Manuskript erscheinen sie zur Hervorhebung in Großbuchstaben.

Beim Versuch, einen irregulären Variablennamen zu vereinbaren, erhalten Sie im SPSS-Dateneditor eine meist informative Fehlermeldung, z.B.:



Tipps zur Benennung:

- Bilden Sie möglichst *informative* Namen, also z.B. FNR, GESCHL und GEBJ für *Fallnummer*, *Geschlecht* und *Geburtsjahr* an Stelle unpraktischer Bezeichnungen wie VAR1, VAR2, VAR3.
- Die eben genannte Regel muss in einem speziellen Fall relativiert werden: Bei Serien verwandter Variablen, z.B. resultierend aus den 12 Fragen des Life Orientation Tests (LOT) im Teil 3 unseres Fragebogens, ist es in der Regel schwer, entsprechend viele individuelle Variablennamen zu bilden. Hier ist meist eine Indexschreibweise günstiger, bei der an einen informativen Namensstamm eine fortlaufende Nummer angehängt wird, z.B. LOT1, LOT2, ...

1.4.3.5 *Kodierplan*

Die Festlegungen zur Strukturierung und Kodierung der Projektdaten sollten in einem **Kodierplan** dokumentiert werden. Er hat zwei Funktionen:

- Während der Erfassung regelt er, wie die Daten eines Falles ins Dateneditorfenster einzutragen bzw. mit einem anderen Programm zu erfassen sind.
- Später dient der Kodierplan als kompakte Beschreibung der entstandenen Datendatei.

Bei unserer KFA-Studie kann für die geplante Erfassung mit dem SPSS-Dateneditor z.B. der folgende Kodierplan verwendet werden:

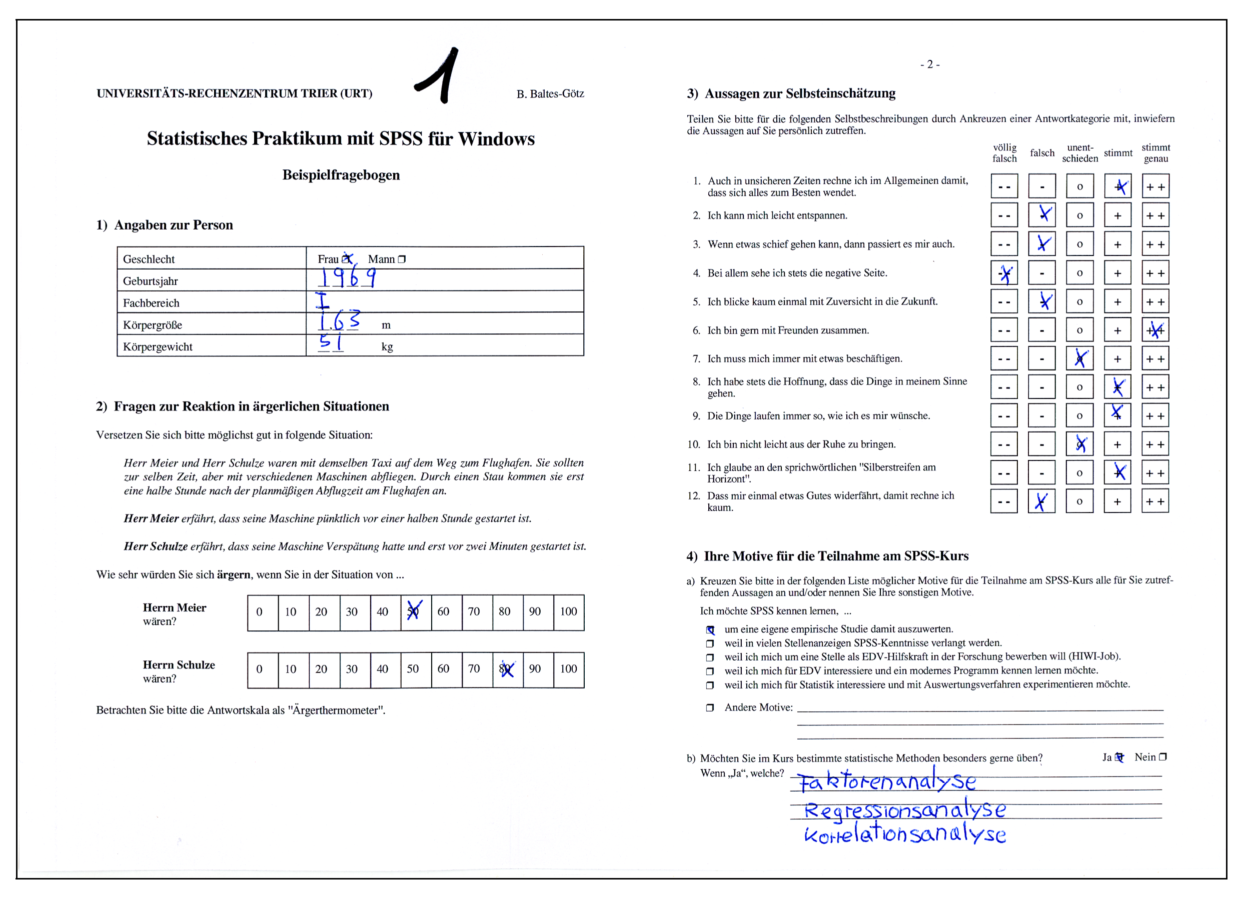
Merkmal	SPSS-Var.-name	Kodierung	Bemerkungen
Fallnummer	FNR	MD-Indikator: entfällt	
Geschlecht	GESCHL	1 = Frau 2 = Mann MD-Indikator: SYSMIS	
Geburtsjahr	GEBJ	**vierstellige** Eingabe (z.B. 1984)! MD-Indikator: SYSMIS	
Fachbereich	FB	1 = I (Pädag., Philos., Psychol.) 2 = II (Sprachen) 3 = III (Hist. und polit. Wiss.) 4 = IV (BWL, Ethnol., Inform., Mathe, Soziol., VWL, Wirtsch.-Inf.) 5 = V (Jura) 6 = VI (Geowissenschaften) 7 = VII (Theologie) MD-Indikator: SYSMIS	
Körpergröße	GROESSE	Eingabe in **cm**! MD-Indikator: SYSMIS	
Körpergewicht	GEWICHT	Eingabe in kg MD-Indikator: SYSMIS	
Ärger als Herr Meier (ohne KFA)	AERGO	0 = 0 1 = 10 . . . 10 = 100 MD-Indikator: SYSMIS	Wir sparen uns per Division durch Zehn viel Schreibarbeit und haben dabei eine zulässige Transformation vorgenommen.
Ärger als Herr Schulze (mit KFA)	AERGM	0 = 0 1 = 10 . . . 10 = 100 MD-Indikator: SYSMIS	
LOT-Fragen	LOT1 bis LOT12	1 = -- 2 = - 3 = o 4 = + 5 = ++ MD-Indikator: SYSMIS	
Kursmotive	MOTIV1 bis MOTIV5, ANDERE	0 = nicht angekreuzt 1 = angekreuzt	SYSMIS wird **nicht** vergeben! Die MD-Behandlung erfolgt später per Datentransformation.
Spezielle Methoden gewünscht?	SMG	0 = nein 1 = ja MD-Indikator: SYSMIS	
Gewünschte statistische Methoden	METH1 bis METH3	1 = Meth.-Kat. 1 gew. 2 = Meth.-Kat. 2 gew. . . . Bei weniger als drei Nennungen: SYSMIS-Initialisierung belassen	Die Kategorienliste wird während der Erfassung nach Bedarf entwickelt und dokumentiert. Die MD-Behandlung erfolgt später per Datentransformation.

Dieser Kodierplan ist bei der Datenerfassung erfreulich einfach zu handhaben und leistet damit einen wichtigen Beitrag zur Integrität der Daten.

Bei der Erfassung mit dem SPSS-Dateneditor (siehe Abschnitt 3.2) werden viele Regeln des Kodierplans in die Variablendeklaration einfließen (vgl. Abschnitt 3.2.2). Dann wird eventuell die Frage auftauchen, ob man nicht auf einen Kodierplan verzichten und das Regelwerk direkt im Deklarationsteil einer SPSS-Datendatei unterbringen kann. Allerdings enthält unser Beispiel viele Vorschriften (z.B. vierstellige Erfassung des Geburtsjahrs, Verlagerung der MD-Behandlung bei den Motiv-Fragen), die per Variablendeklaration nicht hinreichend klar dokumentiert werden können. Bei einem größeren Projekt unter Beteiligung mehrerer Datenerfasser ist ein schriftlicher Kodierplan unbedingt erforderlich. Ist ein Datensatz hingegen zeitlich befristet und nur für eine einzige Person von Interesse, lohnt sich ein Kodierplan nicht.

1.5 Durchführung der Studie (inklusive Datenerhebung)

Bei den obigen Überlegungen zur Strukturierung und Kodierung der Daten hat sich ergeben, dass der in Abschnitt 1.3 wiedergegebene Fragebogen ohne Korrekturen eingesetzt werden kann. Damit steht der Durchführung unserer Befragung nichts mehr im Wege. Im realen Kursverlauf haben die Teilnehmer noch im Zustand der „naiven Unbefangenheit“ (z.B. ohne Kenntnis der KFA-Theorie) die Rolle der Probanden übernommen und so ihre eigenen, von zufälligen Stichprobeneffekten gefärbten Daten produziert. Die Leser(innen) im Selbststudium werden wohl aus praktischen Gründen in der Regel auf die Durchführung einer eigenen KFA-Erhebung verzichten. Im weiteren Verlauf des Manuskripts werden die in einem früheren Kurs erhobenen Daten analysiert. Die zugehörigen Dateien können über das Internet bezogen werden (siehe Vorwort). Hier ist der ausgefüllte Fragebogen derjenigen Untersuchungsteilnehmerin zu sehen, die bei der zufälligen Vergabe einer Fallidentifikation (vgl. Abschnitt 1.4.2.1) die Nummer Eins erhalten hat:



UNIVERSITÄTS-RECHENZENTRUM TRIER (URT) 1 B. Baltes-Götz

Statistisches Praktikum mit SPSS für Windows

Beispielfragebogen

1) Angaben zur Person

Geschlecht	Frau X Mann ❐
Geburtsjahr	1969
Fachbereich	I
Körpergröße	1,63 m
Körpergewicht	51 kg

2) Fragen zur Reaktion in ärgerlichen Situationen

Versetzen Sie sich bitte möglichst gut in folgende Situation:

Herr Meier und Herr Schulze waren mit demselben Taxi auf dem Weg zum Flughafen. Sie sollten zur selben Zeit, aber mit verschiedenen Maschinen abfliegen. Durch einen Stau kommen sie erst eine halbe Stunde nach der planmäßigen Abflugzeit am Flughafen an.

***Herr Meier** erfährt, dass seine Maschine pünktlich vor einer halben Stunde gestartet ist.*

***Herr Schulze** erfährt, dass seine Maschine Verspätung hatte und erst vor zwei Minuten gestartet ist.*

Wie sehr würden Sie sich **ärgern**, wenn Sie in der Situation von ...

Herrn Meier wären?	0	10	20	30	40	50 X	60	70	80	90	100
Herrn Schulze wären?	0	10	20	30	40	50	60	70	80 X	90	100

Betrachten Sie bitte die Antwortskala als "Ärgerthermometer".

- 2 -

3) Aussagen zur Selbsteinschätzung

Teilen Sie bitte für die folgenden Selbstbeschreibungen durch Ankreuzen einer Antwortkategorie mit, inwiefern die Aussagen auf Sie persönlich zutreffen.

	völlig falsch	falsch	unentschieden	stimmt	stimmt genau
1. Auch in unsicheren Zeiten rechne ich im Allgemeinen damit, dass sich alles zum Besten wendet.	- -	-	o	+ X	+ +
2. Ich kann mich leicht entspannen.	- -	- X	o	+	+ +
3. Wenn etwas schief gehen kann, dann passiert es mir auch.	- -	- X	o	+	+ +
4. Bei allem sehe ich stets die negative Seite.	- - X	-	o	+	+ +
5. Ich blicke kaum einmal mit Zuversicht in die Zukunft.	- -	- X	o	+	+ +
6. Ich bin gern mit Freunden zusammen.	- -	-	o	+	+ + X
7. Ich muss mich immer mit etwas beschäftigen.	- -	-	o X	+	+ +
8. Ich habe stets die Hoffnung, dass die Dinge in meinem Sinne gehen.	- -	-	o	+ X	+ +
9. Die Dinge laufen immer so, wie ich es mir wünsche.	- -	-	o	+ X	+ +
10. Ich bin nicht leicht aus der Ruhe zu bringen.	- -	-	o X	+	+ +
11. Ich glaube an den sprichwörtlichen "Silberstreifen am Horizont".	- -	-	o	+ X	+ +
12. Dass mir einmal etwas Gutes widerfährt, damit rechne ich kaum.	- -	- X	o	+	+ +

4) Ihre Motive für die Teilnahme am SPSS-Kurs

a) Kreuzen Sie bitte in der folgenden Liste möglicher Motive für die Teilnahme am SPSS-Kurs alle für Sie zutreffenden Aussagen an und/oder nennen Sie Ihre sonstigen Motive.

Ich möchte SPSS kennen lernen, ...

- [X] um eine eigene empirische Studie damit auszuwerten.
- ❐ weil in vielen Stellenanzeigen SPSS-Kenntnisse verlangt werden.
- ❐ weil ich mich um eine Stelle als EDV-Hilfskraft in der Forschung bewerben will (HIWI-Job).
- ❐ weil ich mich für EDV interessiere und ein modernes Programm kennen lernen möchte.
- ❐ weil ich mich für Statistik interessiere und mit Auswertungsverfahren experimentieren möchte.
- ❐ Andere Motive: ____________________

b) Möchten Sie im Kurs bestimmte statistische Methoden besonders gerne üben? Ja X Nein ❐

Wenn „Ja“, welche? Faktorenanalyse
Regressionsanalyse
Korrelationsanalyse

Diese Nummer wurde nachträglich von der Untersuchungsleitung auf den Fragebogen geschrieben. Bei der in Abschnitt 3.1.1.2 vorzustellenden Daten-Scanner-Lösung *Teleform* kann man die Fallnummern per Seriendruck automatisch anbringen lassen.

2 Einstieg in SPSS für Windows

In den bisher dargestellten Projektphasen von der theoretischen Ausarbeitung bis zur Erstellung des Kodierplans spielte SPSS noch keine wesentliche Rolle. Die im KFA-Projekt nun anstehende Datenerfassung wollen wir jedoch mit diesem Programm bewerkstelligen, so dass an dieser Stelle einige einführende Bemerkungen zu SPSS angemessen sind. In Abschnitt 2.1 geht es um die Verfügbarkeit von SPSS an der Universität Trier, und in den Abschnitten 2.2 bis 2.5 werden elementare Eigenschaften des Programms dargestellt.

2.1 SPSS-Produkte an der Universität Trier

An der Universität Trier steht SPSS für MacOS und Windows mit nahezu identischer Erweiterungsmodul-Ausstattung zur Verfügung:

Regression Models
Advanced Models
Tables
Trends
Categories
Conjoint
Exact Tests (nur Windows)
Missing Values Analysis

Aus der SPSS-Produktfamilie ist außerdem das nur für Windows verfügbare Strukturgleichungsanalyseprogramm Amos vorhanden.

Die aufgeführten SPSS-Produkte können von Angehörigen der Universität Trier im Rahmen ihrer dienstlichen Tätigkeit bzw. ihrer Ausbildung auf folgende Weise genutzt werden:

a) Pool-PCs

Auf den Pool-PCs unter dem Betriebssystem Windows finden Sie über

Start > Alle Programme > Wissenschaftliche Programme

die Programmgruppe **SPSS** mit Unterverzeichnissen zu allen installierten SPSS-Produkten bzw. -Versionen.

b) Kostenlose Nutzung über die URT-Lizenzserver (netzabhängig)

Auf der Webseite

http://www.uni-trier.de/index.php?id=25191

und im Service-Punkt des Rechenzentrums (Eingangsbereich Gebäude E) ist für Angehörige der Universität Trier ein Datenträger samt Installationsanleitung verfügbar. Damit kann SPSS auf einem Rechner unter MacOS oder Windows mit permanentem Internetzugang (an der Uni oder im Privatbereich) zur kostenlosen Nutzung der URT-Lizenzserver installiert werden.

Zur Installation der Programme auf einem Windows-Arbeitsplatzrechner im Campusnetz stehen außerdem automatische Routinen zur Verfügung, die (im Rahmen einer normalen Anmeldung bei der Windows-Domäne URT) über

Start > Systemsteuerung > Software > Neue Programme hinzufügen

erreichbar sind.

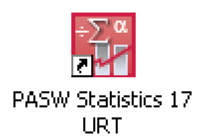
c) Kostenpflichtige individuelle Mietlizenz (netzunabhängig)

Für Rechner unter MacOS oder Windows ohne permanente Netzverbindung zu den URT-Lizenzservern kann in der Benutzerberatung eine befristete SPSS-Einzelplatzlizenz erworben werden.

2.2 Programmstart und Benutzeroberfläche

2.2.1 SPSS starten

Nach erfolgreicher Anmeldung bei einem Pool-PC unter Windows erreichen Sie SPSS 17 über das Desktop-Symbol



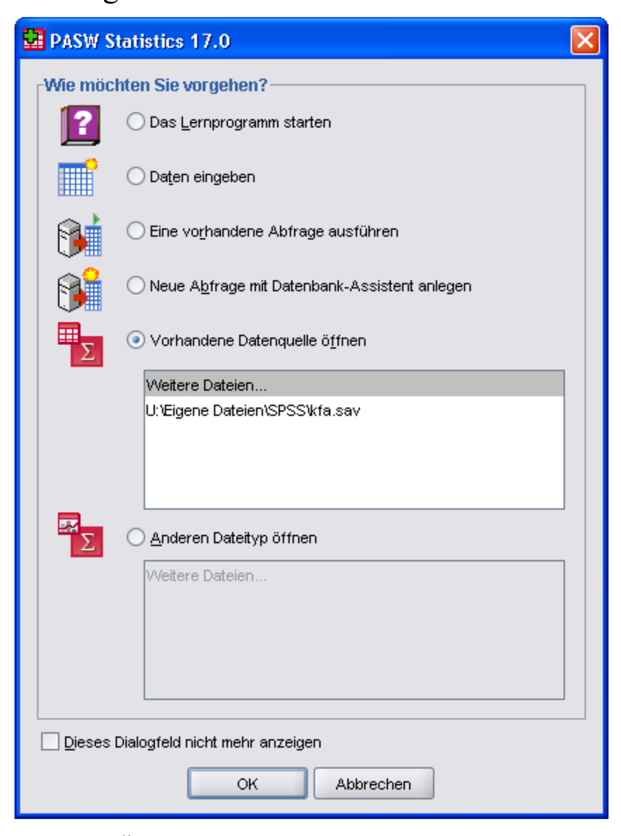
oder über das Startmenü:

Start > Alle Programme > Wissenschaftliche Programme > SPSS > SPSS Statistics 17

Auf einem PC mit lokaler SPSS-Installation können Sie das Programm in der Regel so starten:

Start > Alle Programme > SPSS Inc > SPSS Statistics 17 > SPSS Statistics 17

Nach dem Start erscheint der folgende Assistent:

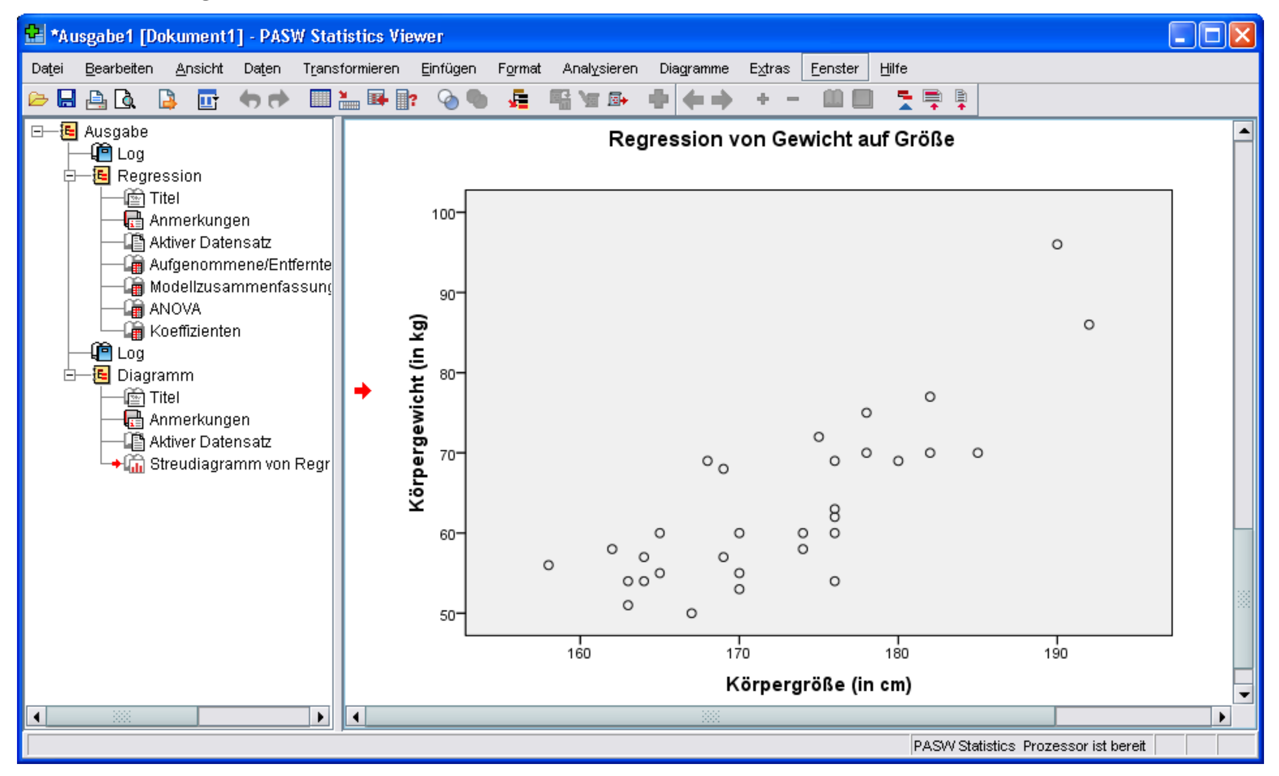


Er ermöglicht z.B. ein bequemes Öffnen der in früheren Sitzungen benutzten Dateien.

2.2.2 Die wichtigsten SPSS-Fenster

Das Dateneditorfenster mit der (Fälle × Variablen) - Datenmatrix haben Sie schon in Abschnitt 1.4.1 kennen gelernt. Nach der Datenerfassung können Sie mit Hilfe seiner Menüzeile statisti-

sche und graphische Datenanalysen anfordern, die dann im **Ausgabefenster**, auch *SPSS Statistics Viewer* genannt, erscheinen, z.B.:



Die SPSS-Fenster enthalten in der Kopfzone eine Menüzeile und verschiebbare Symbolleisten, im Fußbereich eine Statuszeile mit Informationen über wichtige Programmzustände.

2.2.3 Was man mit SPSS so alles machen kann

Wir sind im Moment dabei, einen ersten Eindruck vom Arbeitsplatz *SPSS für Windows* zu gewinnen. Einen guten Überblick vermitteln die Optionen in der Menüzeile des Dateneditorfensters:

- **Datei**
  Hier finden Sie u.a. Befehle zum Öffnen bzw. Sichern von Datendateien sowie zum Beenden von SPSS.
- **Bearbeiten**
  Über das **Bearbeiten**-Menü erreichen Sie Editorbefehle zum Ausschneiden, Kopieren, Einfügen, Löschen und Suchen von Daten sowie die **Optionen**-Dialogbox zur Anpassung von diversen SPSS-Einstellungen. Außerdem können Sie hier Modifikationen des Daten- oder Ausgabefensters rückgängig machen.
- **Ansicht**
  Hier können Sie u.a. die Statuszeile sowie die Symbolleisten aus- bzw. einschalten sowie die Schriftart der angezeigten Daten festlegen.
- **Daten**
  Über das **Daten**-Menü sind Dialoge zur Auswahl einer Teilstichprobe, zur Aggregation von SPSS-Dateien (z.B. mit Daten aus verschiedenen Stichproben) sowie zum Sortieren und Gewichten der Fälle erreichbar.
- **Transformieren**
  Hier finden Sie z.B. die Befehle zum Rekodieren von Variablen oder zum Berechnen neuer Variablen aus bereits vorhandenen.

- **Analysieren**
  Dieser Menüpunkt erschließt die statistischen Auswertungsmethoden, mit denen wir letztlich unsere Forschungsfragen klären wollen.
- **Diagramme**
  An dieser Stelle bietet SPSS vielfältige Möglichkeiten zur graphischen Präsentation von Datenstrukturen an.
- **Extras**
  Hier finden sich diverse Funktionen (z. B. Kommentieren einer Datendatei, Modifikation der SPSS-Menüs).
- **Fenster**
  Über dieses Menü sind die offenen SPSS-Fenster erreichbar.
- **Hilfe**
  Die Hilfefunktion bietet neben systematischen Informationen über das SPSS-System auch ein Lernprogramm, Fallstudien (komplette Anwendungsbeispiele) und einen Statistik-Assistenten.

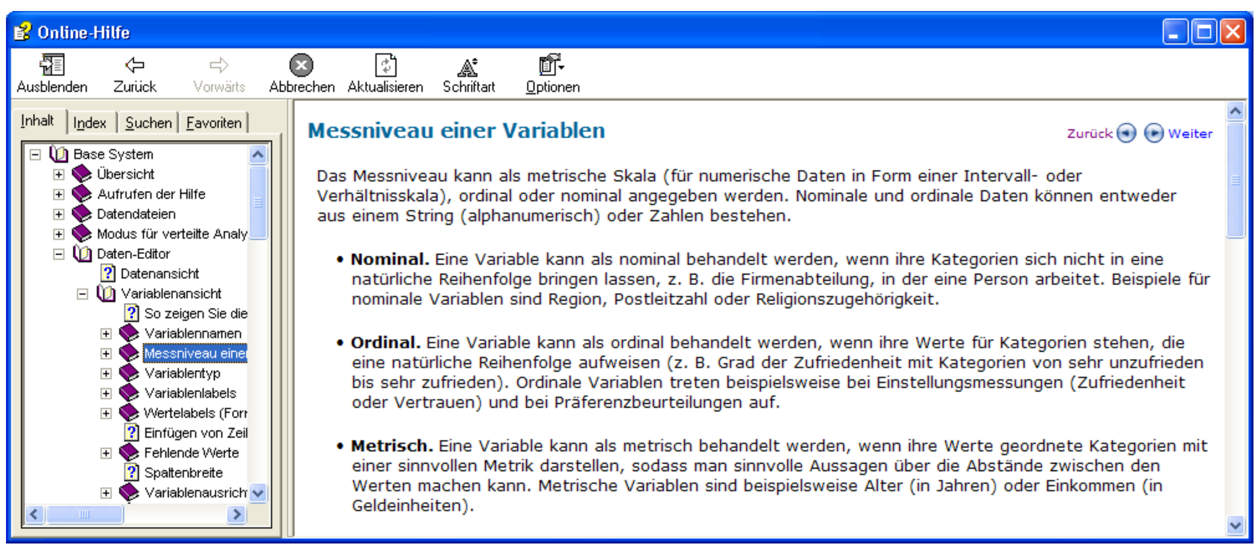
Bei leerem Datenfenster sind die meisten Menüoptionen nicht verfügbar. Die anderen SPSS-Fenster bieten angepasste Menüzeilen.

2.3 Das Hilfesystem

Bei der Arbeit mit SPSS können Sie stets auf ein mächtiges Hilfesystem zurückgreifen, dessen wichtigste Möglichkeiten nun vorgestellt werden.

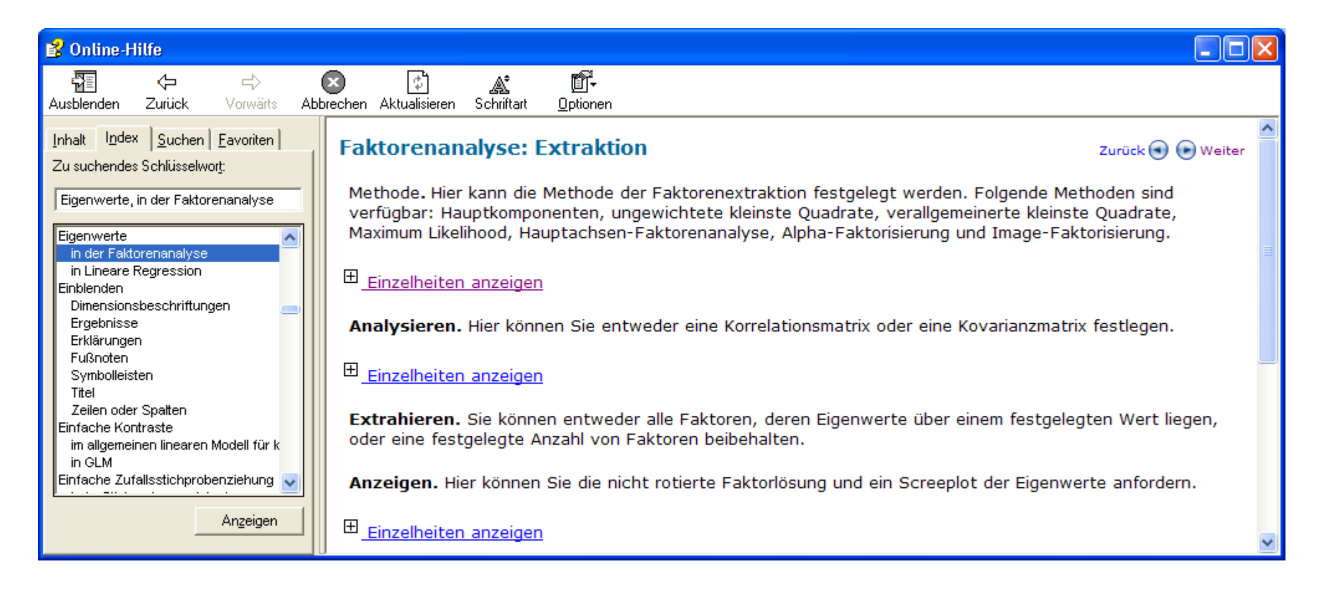
2.3.1 Systematische Informationen

Nach dem Menübefehl **Hilfe > Themen** finden Sie auf der **Inhalt**-Registerkarte des folgenden Fensters Informationen über die installierten SPSS-Erweiterungsmodule in systematischer Form:



2.3.2 Gezielte Suche nach Begriffen

Die Registerblätter **Index** und **Suchen** im Hilfefenster eignen sich für die Suche nach Informationen zu bestimmten Begriffen, z.B.:

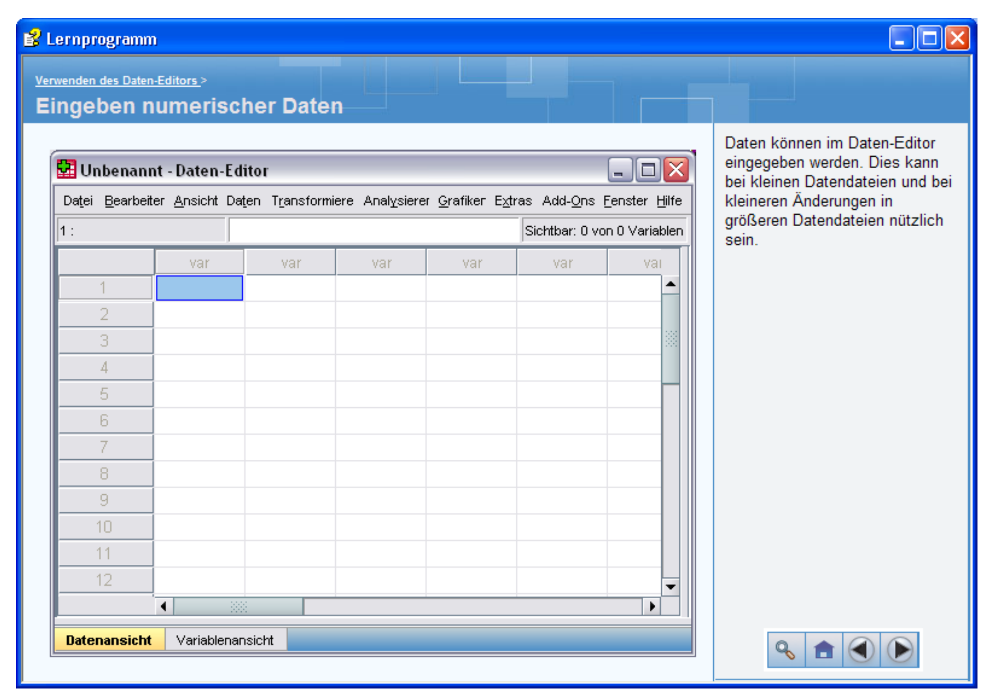


2.3.3 Kontextsensitive Hilfe zu den Dialogboxen

In fast jeder SPSS-Dialogbox können Sie mit der Standardschaltfläche **Hilfe** Informationen zu all ihren Optionen anfordern.

2.3.4 Lernprogramm

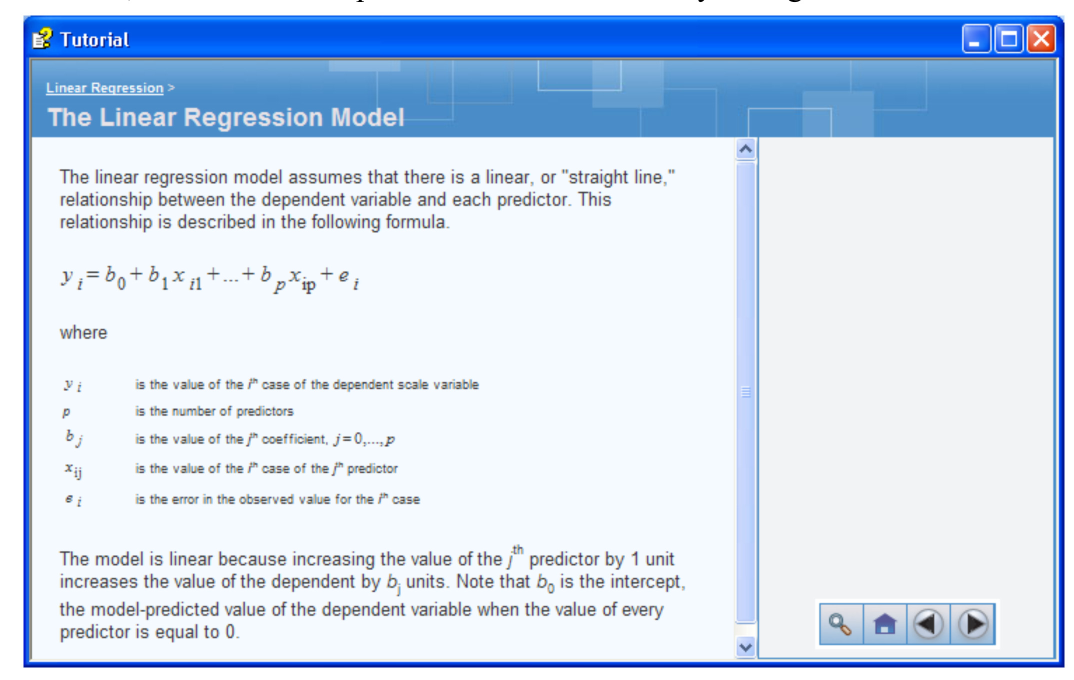
Neben dem eher zum Nachschlagen geeigneten Hilfefenster mit seinen systematischen Beschreibungen und seinem vollständigem Index gibt es ein weiteres Informationsangebot, das eher didaktisch orientiert und daher auf das Wichtigste beschränkt ist: das interaktive SPSS-Lernprogramm. Es wird mit **Hilfe > Lernprogramm** gestartet und sollte mehr oder weniger linear durchgearbeitet werden. In den einzelnen Kapiteln werden konkrete Arbeitsabläufe geübt, z.B.:



Sie können das Lernprogramm als eigenständige Windows-Anwendung parallel zu SPSS ausführen und damit die Lektionen sofort nachvollziehen, indem Sie zwischen SPSS und dem Lernprogramm hin und her wechseln, z.B. mit der Tastenkombination ALT ⇆.

2.3.5 Fallstudien

Nach **Hilfe > Fallstudien** startet ein Tutorium, das mit der interaktiven Technik des Lernprogramms arbeitet, aber den Schwerpunkt auf statistische Analysen legt.



Viele Auswertungsprozeduren werden über ein komplettes Anwendungsbeispiel und Informationen zu folgenden Themen erschlossen:

- Einsatzmöglichkeiten
- Voraussetzungen der Analyse
- Interpretation der Ergebnisse
- Verwandte Verfahren
- Literaturangaben

2.3.6 Statistik-Assistent

Der über **Hilfe > Statistik-Assistent** verfügbare Assistent versucht, den Anwender durch eine Sequenz von Fragen zur richtigen Statistik- bzw. Graphikdialogbox zu führen.

2.4 Weitere Informationsquellen

2.4.1 Handbücher und Manuskripte

Es stehen u.a. zur Auswahl:

- SPSS-Originalhandbücher
  Mit SPSS wird eine umfangreiche Sammlung von PDF-Handbüchern zu den einzelnen Modulen und zu den statistischen Algorithmen ausgeliefert. Allein die Dokumentation der Kommandosprache, über die man die meisten Leistungen des SPSS-Systems abrufen kann (siehe unten), umfasst ca. 2400 Seiten. Dieses PDF-Dokument ist auch im Hilfesystem verfügbar (**Hilfe > Befehlssyntax-Referenz**).
- Sekundärliteratur
  Im Buchhandel und in wissenschaftlich orientierten Bibliotheken finden sich zahlreiche Sekundär-Handbücher zu SPSS. Nach dem Absolvieren des vorliegenden Kurses sind für die meisten SPSS-Anwender(innen) insbesondere diejenigen Bücher von Interesse, wel-

che die jeweils benötigten statistischen Methoden auf einem angemessenen Niveau behandeln und die konkrete Realisation mit SPSS gut unterstützen (z.B. durch eine Erläuterung der Ergebnistabellen). Leider habe ich mir aus Zeitgründen von den zahlreichen Statistik-Lehrbüchern mit SPSS-Unterstützung nur wenige Titel näher ansehen können, so dass die folgende Liste sicher unvollständig ist:

Backhaus et al. (2008). *Multivariate Analysemethoden*

Cohen, et al. (2003). *Applied Multiple Regression/Correlation Analysis …*

Field, A. (2005). *Discovering Statistics Using SPSS*

Norušis (2009). *SPSS 16.0. Statistical Procedures Companion*

Norušis (2009). *SPSS 16.0. Advanced Statistical Procedures Companion*

Tabachnik & Fidell (2007). *Using multivariate statistics*

Die vollständigen bibliographischen Angaben finden Sie im Literaturverzeichnis.

- Auf die URT-Manuskripte zur Verwendung spezieller Analysemethoden in SPSS wurde schon im Vorwort hingewiesen.

2.4.2 SPSS/SPSS im Internet

SPSS bzw. SPSS ist im Internet vielfach präsent, besonders zu erwähnen sind:

- **Die WWW-Homepage der Firma SPSS Inc.**
  http://www.spss.com/
- **Die Usenet-Diskussionsgruppe comp.soft-sys.stat.spss**
  http://groups.google.de/group/comp.soft-sys.stat.spss/topics/
  Hier werden technische und statistische Themen lebhaft diskutiert, wobei SPSS-Mitarbeiter zu wichtigen Fragen kompetent Stellung nehmen.

2.4.3 URT - Service-Punkt

Bei Problemen mit der Anwendung von SPSS-Produkten können sich Angehörige der Universität Trier an den URT - Service-Punkt wenden:

Web:	http://helpdesk.uni-trier.de
Mail:	helpdesk@uni-trier.de
Tel.:	0651-201-4400
Ort:	Foyer Gebäude E
Zeiten:	Montag bis Donnerstag: 07.45-18.15 Uhr, Freitag: 07.45-16.30 Uhr

2.5 SPSS für Windows beenden

Die Beendigung einer SPSS-Sitzung wird mit

Datei > Beenden

eingeleitet. Falls Sie während der Sitzung Dokumente erstellt bzw. verändert und noch nicht gesichert haben (z.B. im Daten- oder im Ausgabefenster), werden Sie von SPSS an das Speichern erinnert.

3 Datenerfassung und SPSS-Dateneditor

Wie bei unserer KFA-Studie liegen auch in vielen anderen Projekten nach Abschluss der Datenerhebung schriftliche Untersuchungsdokumente vor, die nun erfasst, d.h. in eine Computer-Datei übertragen werden müssen. Bevor in Abschnitt 3.2 die konkrete Erfassung der KFA-Daten mit dem SPSS-Dateneditor beschrieben wird, sollen in Abschnitt 3.1 einige alternative Erfassungsmethoden vorgestellt werden.

3.1 Methoden zur Datenerfassung

3.1.1 Automatisierte Verfahren

Zunächst geht es um zwei Optionen zur Rationalisierung der Datenerhebung bzw. -erfassung, die sich zunehmender Beliebtheit erfreuen.

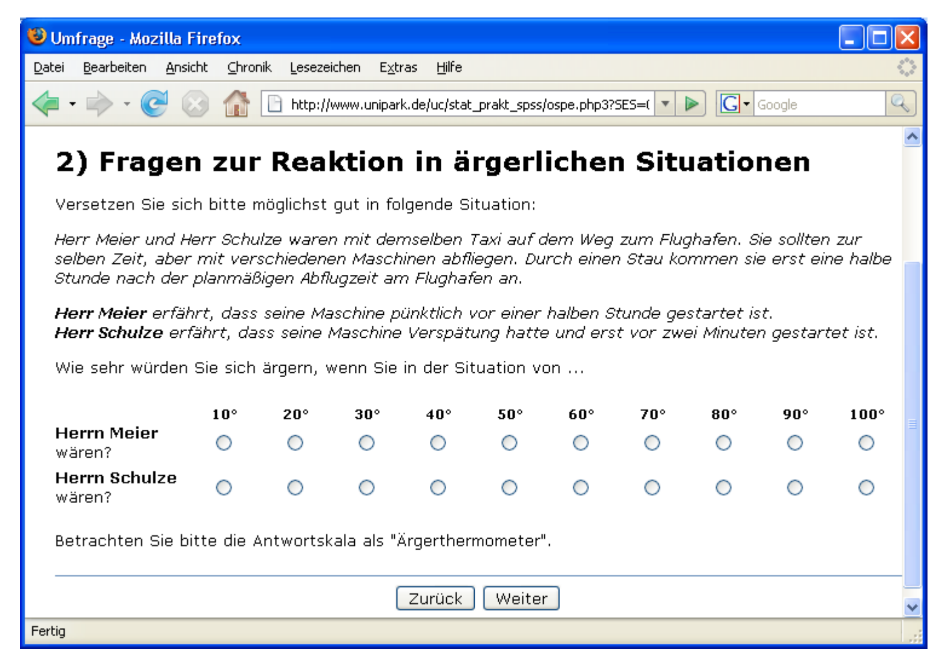
3.1.1.1 Online-Datenerhebung

Wenn die nötigen technischen und organisatorischen Voraussetzungen gegeben sind, sollte eine Online-Datenerhebung eingesetzt werden. Hiermit sind Verfahren gemeint, bei denen die Untersuchungsteilnehmer(innen) ihre Daten (aktiv oder passiv) direkt in eine EDV-Anlage einspeisen (z.B. Internet-Umfrage, automatische Aufzeichnung physiologischer Daten). Nach Abschluss der Datenerhebung kann sofort die Auswertung beginnen, weil die Daten automatisch in einer Datei landen, die oft direkt in SPSS genutzt werden kann. Auf eine gelegentliche Kontrolle (z.B. wegen möglicher Defekte in der Aufzeichnungsapparatur) sollte man aber trotzdem nicht verzichten. Die Daten*erfassung* als eigenständige Arbeitsphase entfällt bei den Online-Verfahren.

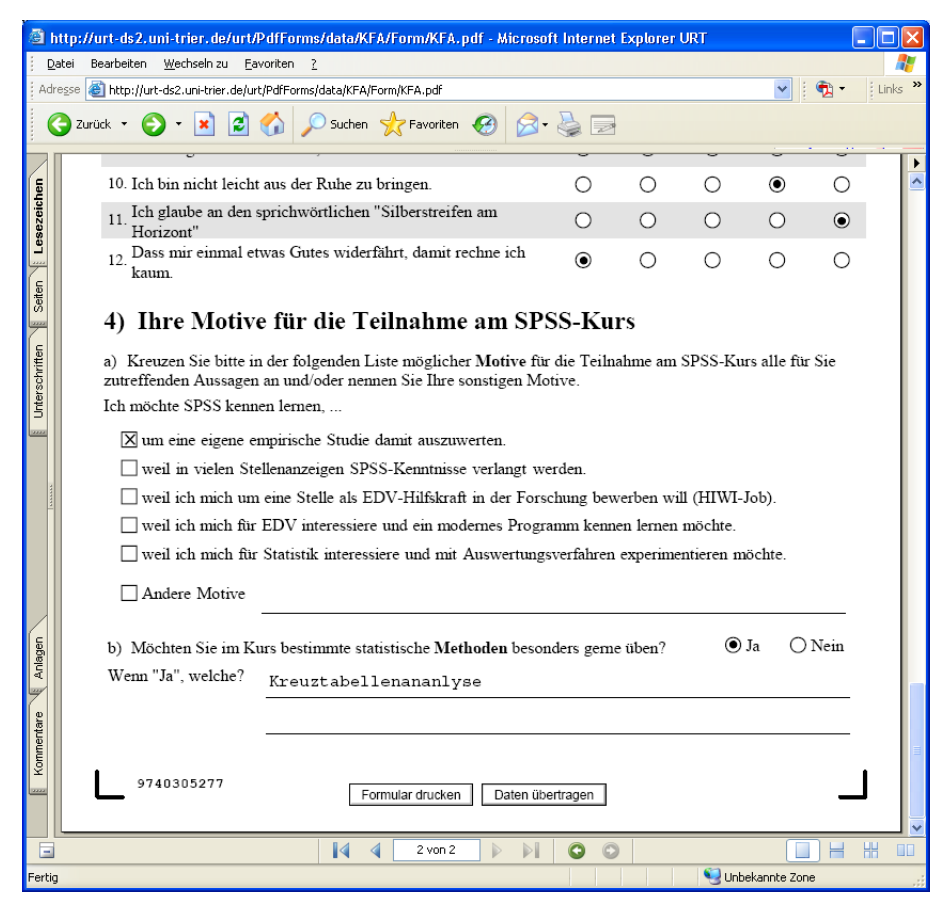
Mit der zunehmenden Verbreitung des Internets verbessern sich Chancen für den Einsatz dieser Kommunikations-Infrastruktur bei einer Vielzahl von Untersuchungen. Allerdings sind u.a. die folgenden Einschränkungen zu beachten:

- Man erreicht (noch) nicht jede Population.
- Für umfangreiche Befragungen ist die Technik weniger geeignet, weil die Unterbrechung und spätere Fortsetzung der anonymen Teilnahme umständlich ist, bei manchen Systemen sogar unmöglich.
- Wenn sich die Online-Umfrageteilnehmer in einer relativ öffentlichen Situation befinden (z.B. PC-Pool einer Hochschule), ist die Auskunftsbereitschaft bei persönlichen Fragen eventuell beschränkt.

Das URT betreibt Online-Umfragesysteme auf HTML- und PDF-Basis (**GlobalPark Enterprise Feedback Suite 7.0**, **Teleform 10.2**), wobei sich z.B. der KFA-Fragebogen mit beiden Systemen gut realisieren lässt. Bei der HTML-basierten GlobalPark-Lösung wird auf Seiten der Umfrageteilnehmer lediglich ein Web-Browser vorausgesetzt:



Bei der auf Teleform (siehe unten) basierenden PDF-Lösung kann das Design des Fragebogens über einen graphischen Editor gestaltet werden. Die Untersuchungsteilnehmer benötigen über den Web-Browser hinaus noch den kostenlos verfügbaren und sehr weit verbreiteten Acrobat-Reader der Firma Adobe:



Wer an der Universität Trier eine Online-Datenerhebung mit GlobalPark oder Teleform durchführen möchte, erhält die erforderliche Zugangsberechtigung und Unterstützung in der Benutzerberatung des Rechenzentrums (vermittelt über den URT - Service-Punkt, siehe Abschnitt 2.4.3).

3.1.1.2 Automatisches Einscannen von schriftlichen Untersuchungsdokumenten

Auch bei einer schriftlichen Befragung im konventionellen Stil lässt sich das manuelle Erfassen der Daten vermeiden. Diese lästige und fehleranfällige Arbeit kann man einer EDV-Anlage zum automatischen Einscannen und Interpretieren der schriftlichen Untersuchungsdokumente übertragen. Allerdings muss die EDV-Anlage erst mit einigem Aufwand in ihre Arbeit eingewiesen werden, so dass bei kleineren Projekten kaum ein Rationalisierungsgewinn zu erzielen ist.

An der Universität Trier steht für die Datenerfassung per Scanner im Graphikraum des Rechenzentrums (E-09) das Programm **Teleform 10.2** mit der erforderlichen Hardware (Scanner mit automatischem Einzelblatteinzug) zur Verfügung. Das Programm kann folgende Informationen erfassen:

- Markierungen der Optionen zu Einfach- oder Mehrfachwahlfragen (OMR)
- gedruckte Zeichen (OCR)
- Handschrift (ICR)

Es enthält einen Formulargenerator, so dass Fragebogendesign und -deklaration in einem Arbeitsschritt erfolgen.

Ein Zusatznutzen besteht in der Möglichkeit, zu einem Teleform-Projekt ein interaktives PDF-Formular mit identischem Design zu erstellen und für eine Online-Umfrage zu verwenden. Damit können Sie entscheiden, ob Sie Ihre Daten

- mit einem gedruckten Fragebogen erheben und per Scanner erfassen,
- per Online-Umfrage erfassen (siehe Abschnitt 3.1.1.1)
- oder parallel über beide Kanäle einsammeln wollen.

Das Teleform-System führt die Daten aus beiden Quellen zusammen und exportiert sie z.B. in eine SPSS-Datendatei.

3.1.2 Manuelle Verfahren

Ist bei einer schriftlichen Datenerhebung der Einsatz einer Scanner-Lösung unrentabel, müssen die Daten per Tastatur unter Beachtung eines Kodierplans in eine Computer-Datei befördert werden. Beim Entwurf des Kodierplans ist darauf zu achten, dass dem Erfasser keine unnötigen und fehleranfälligen Arbeiten zugemutet werden (siehe Abschnitt 1.4).

Von den möglichen manuellen Erfassungsmethoden werden in diesem Manuskript vorgestellt:

- **Erfassung mit dem SPSS-Dateneditor**
  Der SPSS-Dateneditor ist ein integraler Bestandteil des SPSS-Systems, so dass wir uns mit seiner Bedienung auf jeden Fall vertraut machen müssen. Er ist nicht perfekt geeignet für die Erfassung größerer Datenmengen, kann aber in kleinen bis mittleren Projekten verwendet werden. Relativ ähnliche Arbeitsbedingungen für die Datenerfassung bieten Tabellenkalkulationsprogramme wie z.B. MS-Excel.
- **Einsatz eines speziellen Datenerfassungsprogramms**
  Ein spezielles Datenerfassungsprogramm (z.B. SPSS Data Collection) bietet Vorteile gegenüber dem SPSS-Dateneditor, erfordert aber auch zusätzlichen Einarbeitungsaufwand.

Aufgrund des relativ geringen Datenaufkommens in unserem KFA-Projekt ist der SPSS-Dateneditor das optimale Erfassungswerkzeug. Weil in Abschnitt 3.2 die Erfassung der KFA-Daten mit dem SPSS-Dateneditor ausführlich beschrieben wird, können wir uns im aktuellen Abschnitt auf Erläuterungen zu den spezialisierten Datenerfassungsprogrammen beschränken.

Wenn bei *größeren* Projekten eine manuelle Datenerfassung unumgänglich ist, dann sollte in der Regel ein spezielles Datenbankprogramm verwendet werden. Man arbeitet hier bequem mit einer

Erfassungsmaske, die einen *einzelnen* Fall in übersichtlicher Form auf dem Bildschirm präsentiert. Durch folgende Leistungen dieser Spezialprogramme wird die Datenerfassung rationeller und sicherer:

- **Filterfragen** (*Skip & Fill*)
  In Abhängigkeit vom erfassten Wert einer Filtervariablen verzweigen die Datenerfassungsspezialisten zu unterschiedlichen Folgevariablen und versorgen dabei übersprungene Variablen mit einem festgelegten MD-Indikator.
- **Plausibilitätsprüfungen**
  Man kann z.B. dafür sorgen, dass bei der Variablen GESCHL nur die Werte 1, 2 und 9 (als benutzerdefinierter MD-Indikator) akzeptiert werden.

Allerdings entstehen beim Einsatz eines speziellen Datenerfassungsprogramms auch Kosten:

- Es muss ein zusätzliches Programm erlernt werden.
- Für jedes Projekt sind einige Konfigurationsarbeiten erforderlich (z.B. Gestaltung der Erfassungsmaske, Definition der Regeln zur Plausibilitätskontrolle).

Sofern ein Arbeitsplatz mit permanenter Internet-Verbindung zur Verfügung steht, kann auch ein Online-Umfragesystem für die manuelle Dateneingabe mit Erfassungsmaske, Plausibilitätskontrolle und Filterführung eingesetzt werden (vgl. Abschnitt 3.1.1.1). Diese Lösung hat sogar den erheblichen Vorteil, dass an den Erfassungsplätzen als Software nur ein Betriebssystem und ein Web-Browser benötigt werden.

Auch wenn das verwendete Erfassungsprogramm keine SPSS-Datendateien erzeugt, stellt die Übernahme der Daten selten ein Problem dar:

- SPSS unterstützt beim Datenimport zahlreiche Formate (z.B. Textdateien, Excel, SAS, Stata).
- Auf den Pool-PCs der Universität Trier steht mit dem Programm **StatTransfer** ein Konvertierungsspezialist zur Verfügung, der Dateien gängiger Datenbanken oder Statistikprogramme in das SPSS-Format übersetzen kann.

3.2 Erfassung mit dem SPSS-Dateneditor

Für die nächsten Schritte im KFA-Projekt benötigen Sie eine SPSS-Sitzung mit einem leeren Datenfenster. Diese Situation liegt z.B. vor, nachdem Sie SPSS gestartet und den Startassistenten mit dem Ziel **Daten eingeben** verlassen haben. Nötigenfalls können Sie ein leeres Datenfenster mit dem folgenden Menübefehl anfordern:

Datei > Neu > Daten

Im realen SPSS-Kurs steht nun die Variablendeklaration und die Datenerfassung mit dem SPSS-Dateneditor an. Wenn Sie dieses Manuskript im Selbststudium lesen, können und sollten Sie trotzdem die folgenden Arbeitsschritte zur Variablendeklaration konkret nachvollziehen und die Daten des im Manuskript abgedruckten ersten Falles eintragen (siehe Seite 27). Alle Projektphasen nach der Datenerfassung können Sie durch Verwendung der SPSS-Datendatei **kfar.sav** mitmachen, deren Inhalt im weiteren Verlauf erklärt wird. Wie Sie diese Datei von einem Server des Rechenzentrums beziehen können, ist im Vorwort zu erfahren.

3.2.1 Dateneditor, Datenblatt und Arbeitsdatei

SPSS speichert die zu analysierenden Daten während der Sitzung in einer temporären Datei, bezeichnet als **Datenblatt** oder **Daten-Set**. Zur Bearbeitung dient ein **Dateneditorfenster**, das wir der Kürze halber oft als **Datenfenster** bezeichnen. Ein Datenblatt enthält:

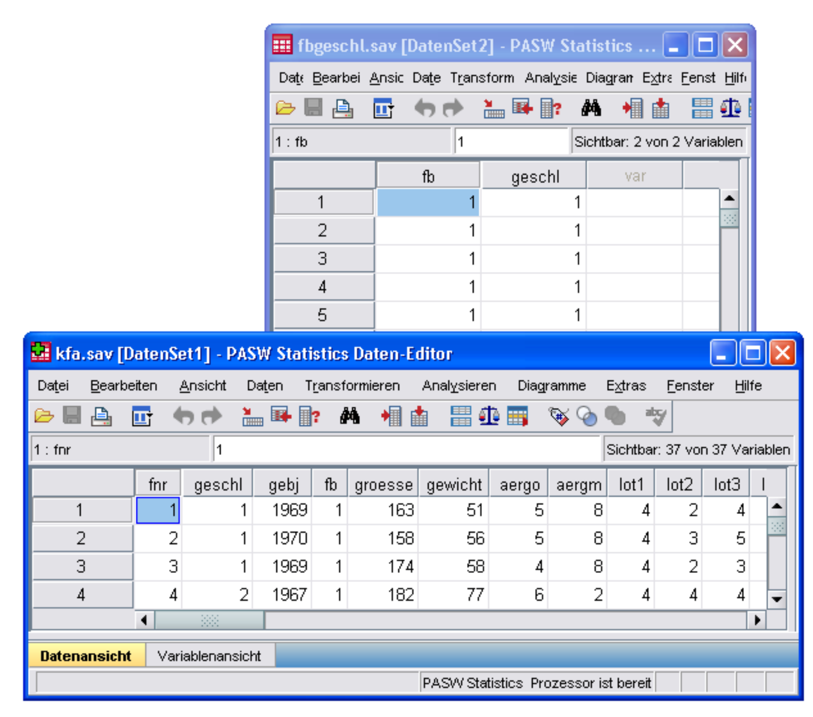
- Die **rechteckige (Fälle × Variablen)-Datenmatrix**
  Diese wird auf dem **Datenansicht**-Registerblatt des Dateneditors bearbeitet.
- Einen so genannten **Deklarationsteil**, auch **Datenlexikon** genannt
  Eine Variable besitzt mehrere verarbeitungsrelevante Attribute, z.B. einen eindeutigen Namen, über den sie bei der Anforderung statistischer oder graphischer Analysen angesprochen werden kann. Zur Verwaltung der Attribute dient das **Variablenansicht**-Registerblatt des Dateneditors.

Mit Hilfe des Dateneditors oder durch Transformationskommandos (siehe unten) können während einer Sitzung u.a. folgende Modifikationen an einem Datenblatt vorgenommen werden:

- Definition von neuen Variablen, Änderung von Variablenattributen (z.B. Namen)
- Manuelle Erfassung von neuen Fällen
  Wir werden in unserem kleinen Demoprojekt die Daten manuell erfassen.
- Löschen von Variablen oder Fällen
- Berechnung neuer Variablen aus bereits vorhandenen.
- Einlesen von Daten aus einer vorhandenen Datei in einem unterstützten Format (z.B. SPSS, Text, MS-Excel, SAS, Stata, ODBC).

Wenn ein Datenblatt über das Ende der Sitzung hinaus erhalten bleiben soll, muss es explizit gesichert werden (in der Regel in eine SPSS-Datendatei, siehe Abschnitt 3.2.5).

Seit der Version 14.0 unterstützt SPSS bzw. SPSS die simultane Verwendung *mehrerer* Dateneditorfenster, z.B.:



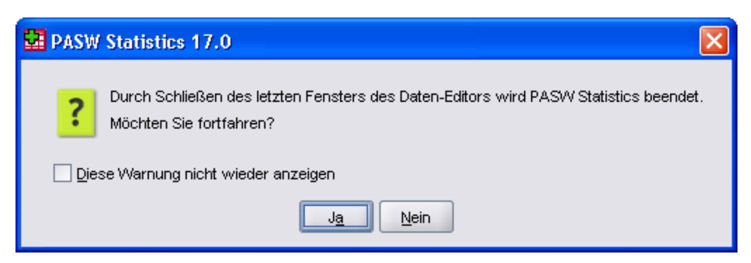
Das Daten-Set zum *aktiven* Dateneditorfenster wird als **Arbeitsdatei** bezeichnet und z.B. bei Analyseanforderungen per Menüsystem angesprochen. Um ein Datenblatt zur Arbeitsdatei zu befördern, muss man das zugehörige Dateneditorfenster per Mausklick oder **Fenster**-Menü in

den Vordergrund holen. Das Datenfenster mit der Arbeitsdatei ist an einem Pluszeichen im Symbol zum Systemmenü zu erkennen (siehe linken Rand der Titelzeile).

Jedes Datenblatt hat einen Namen, welcher in der Titelzeile seines Dateneditorfensters durch eckige Klammern begrenzt hinter dem Namen der verbundenen Datendatei erscheint (siehe oben) und z.B. über den folgenden Menübefehl zu ändern ist:

Datei > Datenblatt umbenennen

Mit dem Schließen des *letzten* Dateneditorfensters beendet man SPSS:



3.2.2 Variablen definieren

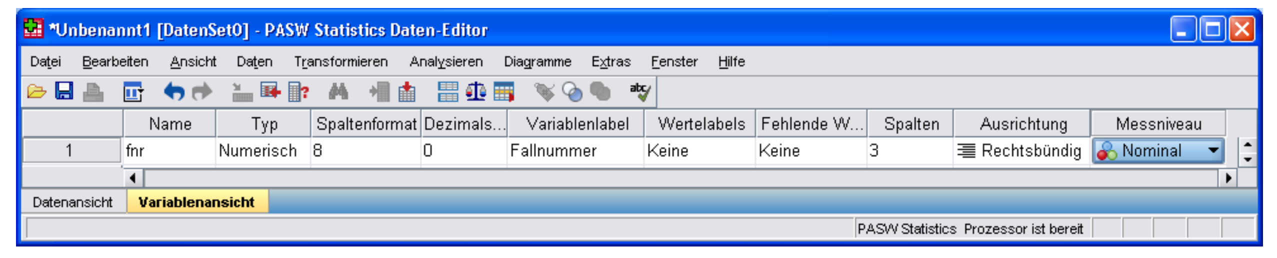
Wie eben erwähnt, verwaltet SPSS für jede Variable eines Datenblatts zahlreiche verarbeitungsrelevante Attribute (z.B. MD-Indikatoren). Diese werden im Deklarationsteil des Datenblatts gespeichert und können vom Anwender festgelegt werden. Da SPSS für alle Attribute geeignete Voreinstellungen benutzt, setzt die Datenerfassung nicht unbedingt eine Variablendefinition voraus[1], doch wird das Erfassen und die spätere Auswertungsarbeit z.B. durch benutzerdefinierte Variablennamen anstelle der automatisch generierten und wenig aussagekräftigen Namen VAR00001, VAR00002 usw. erleichtert. Daher liegt es nahe, dem SPSS-System die in unserem Kodierplan beschriebenen Variablen vor dem Eintragen der Daten bekannt zu machen.

3.2.2.1 Das Datenfenster-Registerblatt Variablenansicht

Ein Datenfenster besitzt *zwei* Registerblätter zur Anzeige bzw. Bearbeitung eines Datenblatts:

- das Registerblatt **Datenansicht** zur Anzeige und Modifikation der (Fälle × Variablen)-Datenmatrix
- das Registerblatt **Variablenansicht** zur Anzeige und Modifikation der Variablenattribute

In einer Zeile der **Variablenansicht** wird jeweils eine Variable beschrieben, wozu in den Spalten insgesamt zehn Attribute zur Verfügung stehen. Für unsere erste Variable (FNR) eignen sich z.B. folgende Angaben:



Um eine neue Variable anzulegen, trägt man ihren Namen in die nächste freie Zeile der Tabelle ein und ändert nach Bedarf die nach dem Verlassen der Namenszelle automatisch generierten

[1] Da in SPSS der Variablentyp *numerisch* voreingestellt ist, müssten wir vor dem Erfassen von Daten anderen Typs auf jeden Fall eine Variablendefinition vornehmen. Allerdings sind solche Variablen in unserem Kodierplan nicht vorgesehen.

Attributvoreinstellungen. Darüber hinaus können auch Variablen eingefügt, gelöscht oder verschoben werden (siehe unten).

3.2.2.2 Die SPSS-Variablenattribute

Bevor wir die Variablen unserer KFA-Studie deklarieren, sollen vorab die SPSS-Variablenattribute erläutert werden:

- **Name**
  Die wesentlichen Regeln für SPSS-Variablennamen wurden schon im Zusammenhang mit dem Kodierplan genannt (siehe Seite 25).
- **Typ**
  Die wichtigsten SPSS-Variablentypen sind schon benannt: Numerisch, String und Datum (siehe Seite 20). In der Regel empfiehlt es sich, auch bei nominalskalierten Merkmalen eine numerische Kodierung vorzunehmen (siehe Abschnitt 1.4.3), so dass der voreingestellte numerische Variablentyp meist beibehalten werden kann.
- **Spaltenformat**
  Bei einer *numerischen* Variablen beeinflusst dieses Attribut lediglich ihre voreingestellte Breite bei der Ausgabe in eine Textdatendatei über das Kommando WRITE (inkl. Vorzeichen und Dezimaltrennzeichen) und ist daher für die Arbeit mit dem Daten- und dem Ausgabefenster wenig relevant. Allerdings muss der Spaltenformatwert stets größer sein als die Anzahl der Dezimalstellen (siehe unten).

  Bei einer *alphanumerischen* Variablen legt das Spaltenformat die maximale Anzahl der gespeicherten Zeichen fest und ist folglich recht bedeutsam. So werden z.B. bei einer nachträglichen Reduktion der Spaltenzahl tatsächlich entsprechend viele Zeichen am rechten Rand gelöscht.
- **Dezimalstellen**
  Bei einer *numerischen* Variablen können Sie festlegen, welche Anzahl von Dezimalstellen bei der Anzeige ihrer Werte im Datenfenster bzw. in der Ergebnisausgabe verwendet werden soll. Diese Angabe betrifft *nicht* die Speichergenauigkeit, sondern nur die Anzeige. Bei einer *alphanumerischen* Variablen ist das Attribut irrelevant und auf den Wert Null fixiert.
- **Variablenlabel**
  Hier können optional Variablenlabel mit einer maximalen Länge von 256 Zeichen vereinbart werden, die in Ergebnistabellen und Graphiken an Stelle der aus praktischen Erwägungen möglichst kurz gewählten und mit Syntaxrestriktionen belasteten Variablennamen (z.B. Verbot von Leer- und Sonderzeichen) angezeigt werden sollen, z.B.:

Variablenname	Variablenlabel
FB	Fachbereich an der Universität Trier
GEWICHT	Körpergewicht (in kg)

  Allerdings erscheinen die Labels in der Ausgabe mancher SPSS-Prozeduren nicht in voller Länge.

  Während wir die Variablen*namen* in SPSS der Einfachheit halber stets klein schreiben, ist bei den Variablen*labels* eine publikationsreife Groß/Kleinschreibung angemessen.
- **Wertelabels**
  Hier können optional Wertelabels mit maximal 60 Zeichen zur Erläuterung von Variablenausprägungen vereinbart werden, was speziell bei numerisch kodierten nominalskalierten Merkmalen empfehlenswert ist, z.B.:

Variablenname	Werte	Wertelabels
GESCHL	1 2	Frau Mann

Im nächsten Abschnitt wird die Definition von GESCHL-Wertelabels konkret vorgeführt.

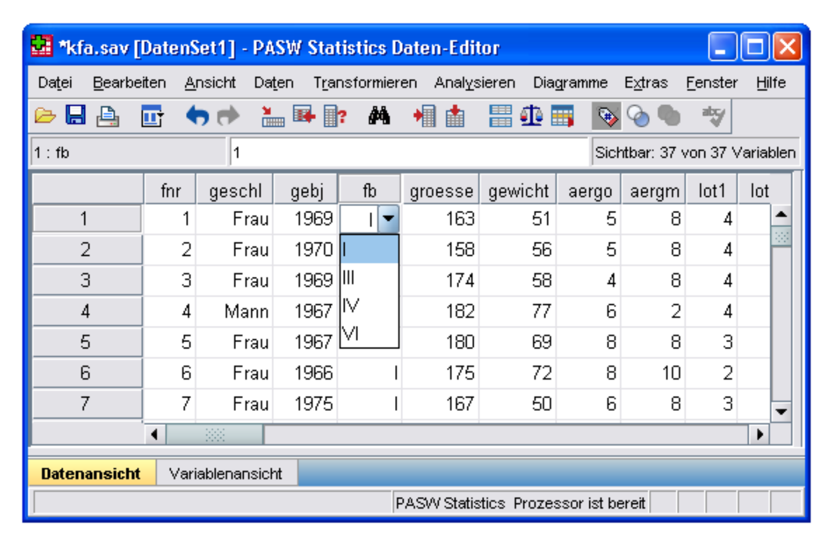
Kommt bei einer nominalen oder ordinalen Variablen die Beteiligung bei einer Diagrammerstellung (siehe Abschnitt 9.1.1) in Frage, sollten auf jeden Fall Wertelabels vergeben werden. Man erhält ansehnliche Beschriftungen (z.B. von Balken) und beeinflusst auch die Berücksichtigung von Kategorien: Soll eine unbesetzte Kategorie in einer Graphik erscheinen (z.B. als Balken mit der Höhe Null), muss ein Wertelabel vergeben werden.

In der Datenansicht bietet der Dateneditor über den Menübefehl

Ansicht > Wertelabels

bzw. den Symbolschalter einige Unterstützung für die Etiketten:

- Sie werden an Stelle der Werte angezeigt.
- Alternativ zur Werteingabe per Tastatur kann man per Drop-Down-Menü ein Label wählen:



Viele SPSS-Anwender(innen) überschätzen allerdings die Rolle der Wertelabels bei der Datenerfassung: Es ist z.B. *nicht* möglich, durch Vergabe von Wertelabels die Menge der gültigen Werte einer Variablen zu definieren und eine Plausibilitätskontrolle für die Erfassung per Dateneditor einzurichten. Trotz obiger Wertelabels-Vereinbarung wird der SPSS-Dateneditor bei der Variablen GESCHL beliebige Zahlen akzeptieren.

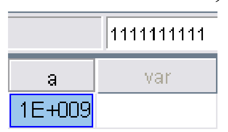
- **Fehlende Werte**

  Wenn Sie bei einer Variablen *benutzerdefinierte* MD-Indikatoren verwenden wollen, müssen Sie diese unbedingt deklarieren, weil sie sonst wie gültige Werte behandelt werden, z.B. bei einer Mittelwertsbildung. Da wir im KFA-Projekt laut Kodierplan ausschließlich System-Missing als MD-Indikator verwenden, müssen wir anschließend keine MD-Deklaration vornehmen (vgl. Abschnitt 1.4.3.5). Daher wird nun die simple Prozedur zum Deklarieren von benutzerdefinierten MD-Indikatoren beschrieben:

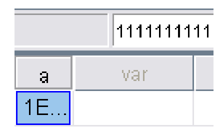
  - Markieren Sie bei der betroffenen Variablen die Zelle zum Attribut **Fehlende Werte**.
  - Nach einem Mausklick auf den nun vorhandenen Erweiterungsschalter ... erscheint eine Dialogbox, in der man entweder bis zu drei Einzelwerte oder aber ein Intervall samt zusätzlichem Einzelwert als MD-Indikatoren vereinbaren kann, z.B.:



- **Spalten** und **Ausrichtung**
  Wie breit soll die Spalte einer Variablen im Dateneditorfenster sein? Wie sollen die Werte ausgerichtet werden (linksbündig, zentriert, rechtsbündig)? Die Attribute **Spalten** und **Ausrichtung** wirken sich nur auf die Darstellung einer Variablen im Datenfenster aus. Genügt die **Spalten**-Zahl nicht, versucht SPSS bei numerischen Variablen, das Platzproblem durch die Exponentialschreibweise zu lösen, z.B.:



  Noch größere Platznot signalisiert SPSS durch Pünktchen:



- **Messniveau**
  Über die technischen Variablenattribute hinaus kann das Messniveau der Variablen deklariert werden, wobei die Ausprägungen *metrisch*, *ordinal* und *nominal* möglich sind. Bisher spielt das deklarierte Messniveau der Variablen bei den meisten SPSS-Prozeduren noch keine Rolle. Bei der Diagrammerstellung (siehe Abschnitt 9.1.1) hängt die Behandlung einer Variablen jedoch vom deklarierten Messniveau ab, und SPSS empfiehlt ausdrücklich, für alle bei einem Diagramm beteiligten Variablen das Messniveau korrekt anzugeben.

3.2.2.3 Variablendefinition durchführen

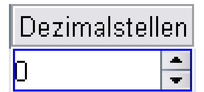
Aktivieren Sie nun die **Variablenansicht** des Datenfensters, und tragen Sie für die erste Variable (zur Fallidentifikation) den Namen FNR ein. Nach dem Markieren der zugehörigen Zelle können Sie sofort mit dem Eintippen des Namens beginnen. Die Groß/Kleinschreibung ist bzgl. der Identifikation von Variablen irrelevant. Die folgenden Namen bezeichnen allesamt dieselbe Variable:

fnr Fnr FNR

Im Manuskript werden Variablennamen aus darstellungstechnischen Gründen groß geschrieben.

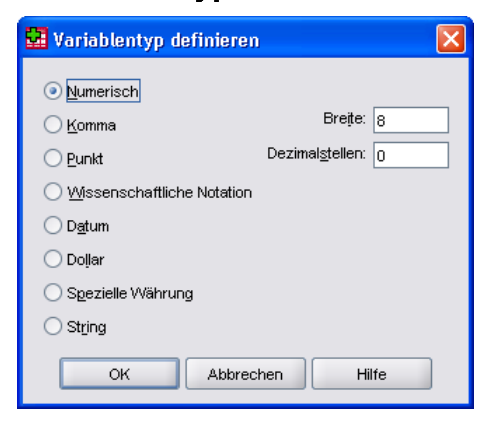
Sobald Sie die Zelle mit dem Variablennamen verlassen (z.B. per Mausklick auf eine andere Zelle oder per Tabulatortaste) wird eine neue Variable mit dem gewünschten Namen in die Arbeitsdatei aufgenommen, sofern gegen den Variablennamen keine Einwände bestehen, und die restlichen Attribute der neuen Variablen werden mit Standardwerten versorgt.

Nach dem Markieren der Zelle **Dezimalstellen** kann man die gewünschte Anzahl von Dezimalstellen durch Eingabe einer Zahl oder per Up-Down - Regler wählen:

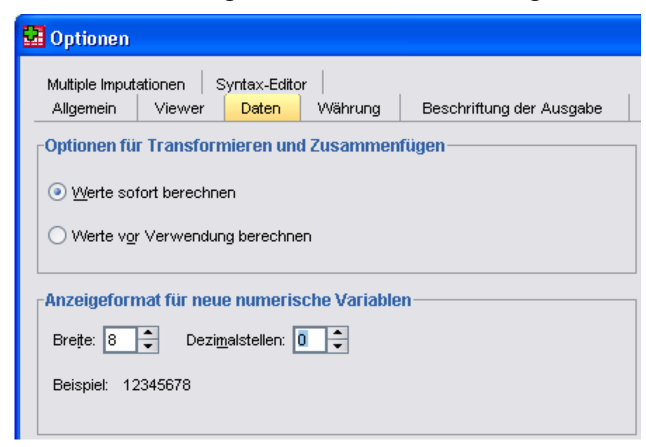


Analog wird auch das Attribut **Spaltenformat** festgelegt, das allerdings bei der von uns geplanten Arbeitsweise keine große Rolle spielt (siehe oben).

Eine alternative Möglichkeit zum Einstellen der Attribute **Dezimalstellen** und **Spaltenformat** findet sich in der Dialogbox **Variablentyp definieren**, die nach einen Mausklick auf den Erweiterungsschalter ... in der markierten **Typ**-Zelle erscheint:

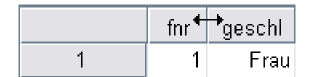


Tipp: Wenn in einem Projekt das voreingestellte Anzeigeformat für numerische Variablen (Breite = 8, Dezimalstellen = 2) häufig durch eine bestimmte Alternative ersetzt werden muss, kann zur Vereinfachung der Deklaration die Voreinstellung entsprechend geändert werden. Dazu öffnet man mit **Bearbeiten > Optionen** die Dialogbox **Optionen**, wechselt hier zum Registerblatt **Daten** und nimmt im Rahmen **Anzeigeformat für neue numerische Variablen** die gewünschten Einstellungen vor, z.B.:



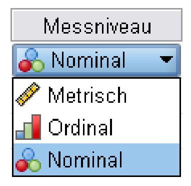
Wenngleich die Variable FNR im Ausgabefenster nicht allzu oft auftauchen wird, tragen wir in die Zelle zum Attribut **Variablenlabel** den Text *Fallnummer* ein.

Statt die Breite der FNR-Spalte im Datenfenster über eine gut geschätzte **Spalten**-Angabe festzulegen, können Sie bei aktiviertem Registerblatt **Datenansicht** auch folgendermaßen vorgehen: Setzen Sie den Mauszeiger auf den rechten Rand der Zelle mit dem Variablennamen, woraufhin der Zeiger eine neue Form und dementsprechend eine neue Funktion erhält:

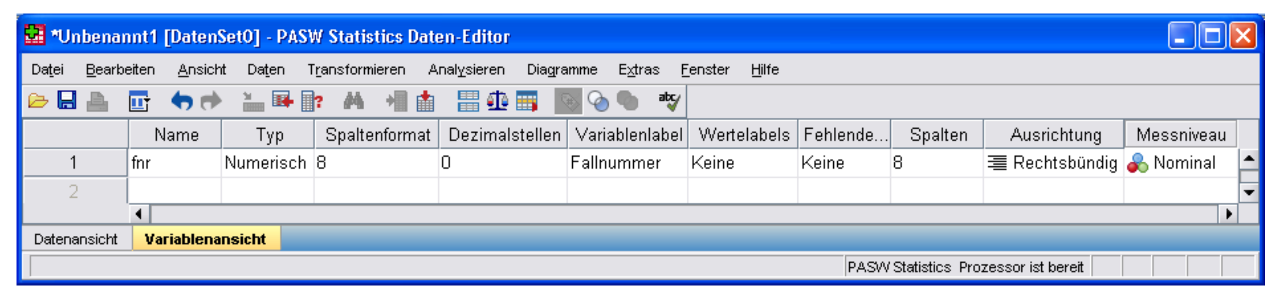


Nun lässt sich der rechten Rand der aktuellen Spalte verschieben: Linke Maustaste drücken, ziehen und an der gewünschten Position wieder loslassen. Eine so festgelegte Spaltenbreite wird von SPSS als **Spalten**-Wert übernommen.

Öffnen Sie in der markierten **Messniveau**-Zelle das Drop-Down-Menü, um für die Fallnummer ein nominales Skalenniveau zu deklarieren:



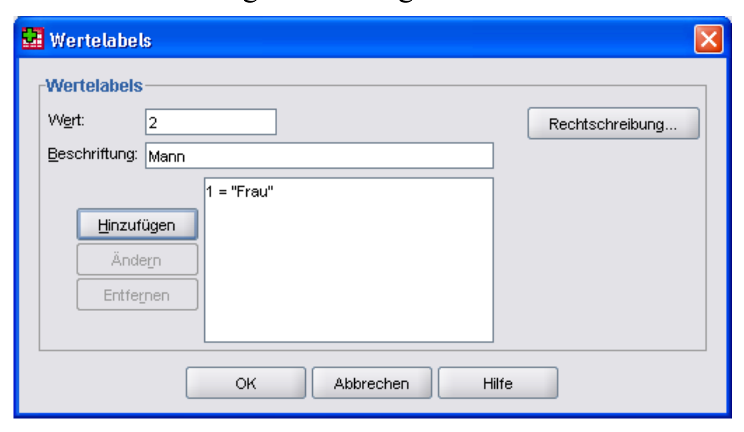
MD-Indikatoren müssen wir (dank SYSMIS-Option) im KFA-Projekt generell nicht vereinbaren (vgl. Abschnitt 1.4.3.2.2), Wertelabels sind bei der augenblicklich bearbeiteten Fallnummernvariablen irrelevant, und das Attribut **Ausrichtung** übernehmen wir stets unverändert. Daher können wir die Deklaration der Variablen FNR beenden:



Bei Bedarf sind Anpassungen jederzeit möglich.

Vereinbaren Sie nun in der zweiten Zeile der Variablenansicht für die Geschlechtsvariable den Namen GESCHL, eine Anzeige ohne Dezimalstellen und das Variablenlabel *Geschlecht*.

Bei diesem numerisch kodierten nominalskalierten Merkmal ist es sehr empfehlenswert, die willkürliche Zuweisung von Zahlen zu den beiden Kategorien durch Wertelabels zu dokumentieren, damit wir bei der Lektüre von Ergebnisausgaben nicht rätseln müssen, welches Geschlecht die Nummer Eins ist. Öffnen Sie daher mit einem Mausklick auf den Erweiterungsschalter [...] in der markierten **Wertelabels**-Zelle die folgende Dialogbox:



Legen Sie z.B. das weibliche Label folgendermaßen fest:

- Tragen Sie den **Wert** 1 und das **Wertelabel** *Frau* ein.
- Drücken Sie auf den Schalter **Hinzufügen**.

  In der Schaltflächen-Beschriftung **<u>H</u>inzufügen** signalisiert das unterstrichene **<u>H</u>**, dass der Mausklick auf die Schaltfläche durch die Tastenkombination **Alt+H** ersetzt werden kann.

Abschließend ist für GESCHL noch das nominale **Messniveau** zu deklarieren.

3.2.2.4 Übung

Definieren Sie alle Variablen zur ersten Seite unseres KFA-Fragebogens. Wie Sie nötigenfalls Variablen einfügen oder löschen können, erfahren Sie im nächsten Abschnitt.

3.2.3 Variablen einfügen, löschen oder verschieben

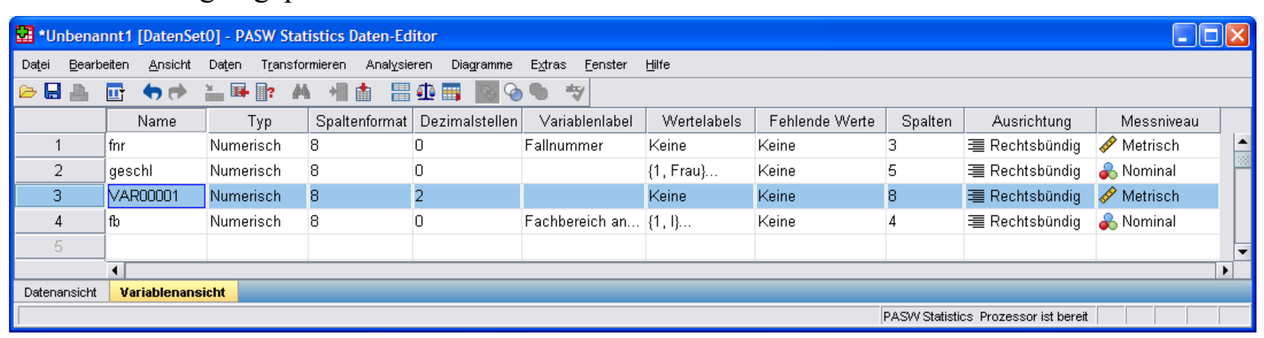
Bei der Variablendefinition kann sich leicht die Notwendigkeit ergeben, Variablen einzufügen oder zu löschen.

3.2.3.1 Variablen einfügen

Wenn Sie z.B. nach FNR und GESCHL die Variable FB definiert und folglich die Variable GEBJ vergessen haben, können Sie das Missgeschick in der Variablenansicht folgendermaßen korrigieren:

- Setzen Sie einen rechten Mausklick auf die Nummer der FB-Zeile (am linken Rand der Tabelle).
- Wählen Sie aus dem Kontextmenü die Option **Variable einfügen**.

Daraufhin stellt SPSS vor FB eine neue Variable mit voreingestellten Attributen zur Verfügung, die nun beliebig angepasst werden können:



Auf analoge Weise lässt sich eine neue Variable auch in der Datenansicht einfügen:

- Setzen Sie einen rechten Mausklick auf die Beschriftung der FB-Spalte im Kopfbereich der Tabelle.
- Wählen Sie die Option **Variablen einfügen** aus dem Kontextmenü.

3.2.3.2 Variablen löschen

Gehen Sie in der Variablenansicht folgendermaßen vor, um eine Variable zu löschen:

- Setzen Sie einen rechten Mausklick auf die Zeilennummer der betroffenen Variablen (am linken Rand der Tabelle).
- Wählen Sie die aus dem Kontextmenü Option **Löschen**.

Auf analoge Weise lässt sich eine Variable auch in der Datenansicht löschen.

3.2.3.3 Variablen verschieben

Gehen Sie in der Variablenansicht folgendermaßen vor, um Variablen per Drag & Drop (Ziehen und Ablegen) zu verschieben:

- Markieren Sie die zu verschiebenden Variablen auf Windows-übliche Weise über Maus-aktionen im Nummerierungsbereich, ggf. ergänzt durch die **Strg**- oder Umschalt-Taste. Lassen Sie anschließend die Maustaste wieder los.

- Klicken Sie in der Nummerierungszone erneut auf die zu verschiebende Variablenauswahl, und halten Sie dabei die linke Maustaste gedrückt.
- Bewegen Sie bei gedrückter Maustaste den Mauszeiger zum Ziel der Verschiebungsaktion. Der aktuell anvisierte Zielort wird von SPSS durch eine rote Linie gekennzeichnet.
- Wenn Sie die Maustaste loslassen, erscheinen die Variablen am neuen Ort.

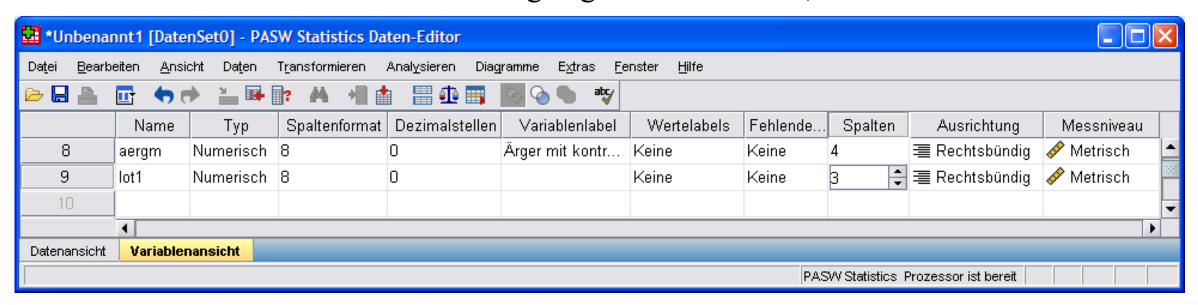
Auf analoge Weise lässt sich eine Variablenauswahl auch in der Datenansicht verschieben.

3.2.4 Attribute auf andere Variablen übertragen

3.2.4.1 Variablendeklarationen vervielfältigen

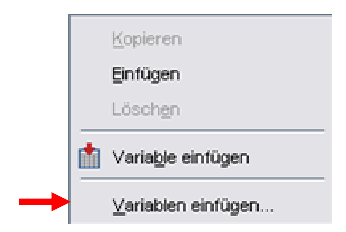
Für unsere zwölf LOT-Fragen sollen natürlich alle Variablenattribute mit Ausnahme des Namens identisch sein. Erfreulicherweise müssen wir die identische Variablendefinition nicht 12-mal wiederholen, sondern können nach einer ersten Definition die Attribute auf alle anderen Variablen übertragen. Mit der folgenden Vorgehensweise lässt sich sogar das Schreiben der restlichen Variablennamen automatisieren:

- Deklarieren Sie die Variable LOT1 mit geeigneten Attributen, z.B.:



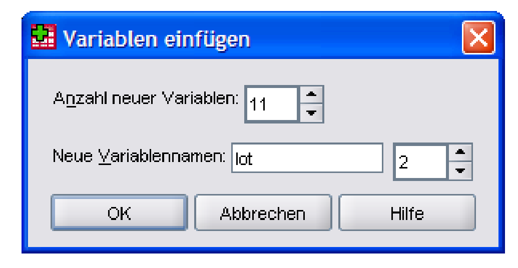
  Das voreingestellte metrische Messniveau kann beibehalten werden, obwohl die fünfstufigen Variablen LOT1 bis LOT12 wohl eher grobschlächtige Indikatoren für das angenommene latente Merkmal Optimismus sind. In den geplanten Auswertungen werden wir nicht die zwölf Rohvariablen selbst, sondern eine daraus abgeleitete Mittelwertsvariable verwenden, für die ein approximativ metrisches Messniveau angenommen werden darf.

- Öffnen Sie das Kontextmenü zur Variablen LOT1 per Mausrechtsklick auf ihre Zeilennummer am linken Tabellenrand, und wählen Sie das Item **Kopieren**, um alle Attribute der Variablen in die Winddows-Zwischenablage zu befördern.
- Setzen Sie einen rechten Mausklick auf die Nummer der nächsten freie Zeile der Variablenansicht und wählen Sie aus dem Kontextmenü die Option **Variablen einfügen** mit den drei Punkten am Ende der Beschriftung:



  Diese Option ist nur verfügbar, wenn sich eine komplette Variablenbeschreibung in der Zwischenablage befindet.

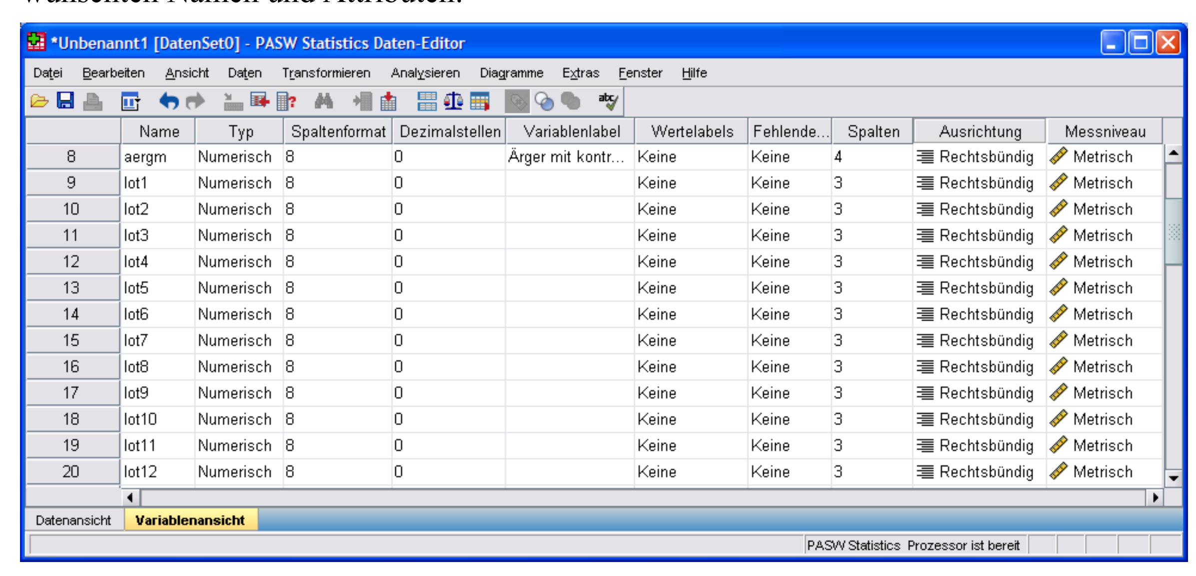
- In der folgenden Dialogbox



können Sie nun festlegen, …

  - wie viele neue Variablen benötigt werden,
  - welche gemeinsame Wurzel die neuen Variablennamen haben sollen,
  - mit welchem Indexwert SPSS den Namen der *ersten* Variablen komplettieren soll.

Nach dem Quittieren der obigen Dialogbox entstehen elf neue Variablen mit den gewünschten Namen und Attributen:



	Name	Typ	Spaltenformat	Dezimalstellen	Variablenlabel	Wertelabels	Fehlende...	Spalten	Ausrichtung	Messniveau
8	aergm	Numerisch	8	0	Ärger mit kontr...	Keine	Keine	4	Rechtsbündig	Metrisch
9	lot1	Numerisch	8	0		Keine	Keine	3	Rechtsbündig	Metrisch
10	lot2	Numerisch	8	0		Keine	Keine	3	Rechtsbündig	Metrisch
11	lot3	Numerisch	8	0		Keine	Keine	3	Rechtsbündig	Metrisch
12	lot4	Numerisch	8	0		Keine	Keine	3	Rechtsbündig	Metrisch
13	lot5	Numerisch	8	0		Keine	Keine	3	Rechtsbündig	Metrisch
14	lot6	Numerisch	8	0		Keine	Keine	3	Rechtsbündig	Metrisch
15	lot7	Numerisch	8	0		Keine	Keine	3	Rechtsbündig	Metrisch
16	lot8	Numerisch	8	0		Keine	Keine	3	Rechtsbündig	Metrisch
17	lot9	Numerisch	8	0		Keine	Keine	3	Rechtsbündig	Metrisch
18	lot10	Numerisch	8	0		Keine	Keine	3	Rechtsbündig	Metrisch
19	lot11	Numerisch	8	0		Keine	Keine	3	Rechtsbündig	Metrisch
20	lot12	Numerisch	8	0		Keine	Keine	3	Rechtsbündig	Metrisch

3.2.4.2 *Alle Attribute einer Variablen auf andere Variablen übertragen*

Gehen Sie z.B. folgendermaßen vor, um alle Attribute einer Variablen (mit Ausnahme des Namens) auf andere, bereits vorhandene Variablen zu übertragen:

- Öffnen Sie das Kontextmenü zur Quellvariablen per Mausrechtsklick auf ihre Zeilennummer am linken Tabellenrand, und wählen Sie das Item **Kopieren**, um alle Attribute der Variablen in die Zwischenablage zu befördern.
- Markieren Sie *eine* Zielvariable per Mausklick auf ihre Zeilennummer oder eine Serie von Zielvariablen durch Mausklicks in Kombination mit der Umschalt- oder **Strg**-Taste.
- Übertragen Sie die in der Zwischenablage gespeicherten Attribute über das Kontextmenü-Item **Einfügen**, mit der Tastenkombination **Strg+V** oder mit dem Menübefehl

  Bearbeiten > Einfügen

  auf alle markierten Variablen.

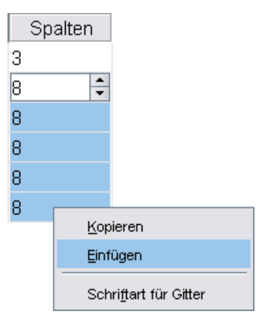
3.2.4.3 *Einzelne Attribute einer Variablen auf andere Variablen übertragen*

Es ist auch möglich, ein *einzelnes* Attribut von einer Variablen auf andere zu übertragen:

- Kontextmenü zur Quell-Attributzelle durch einen rechten Mausklick öffnen und das Item **Kopieren** wählen, z.B.:



- Zu verändernde Attributzellen mit der rechten Maustaste markieren



und den Attributwert über das Kontextmenü-Item **Einfügen** aus der Zwischenablage übernehmen.

3.2.4.4 Übung

Definieren Sie die restlichen Variablen unserer KFA-Studie.

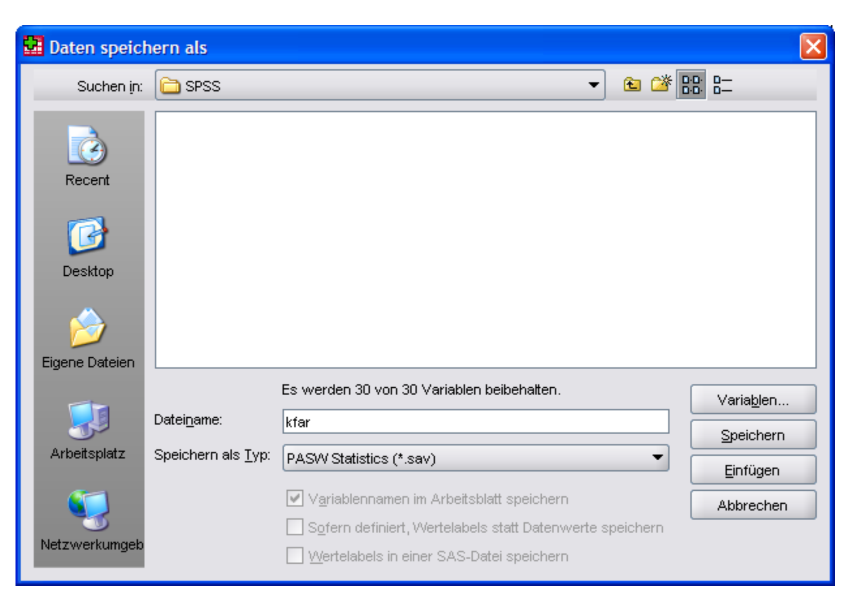
3.2.5 Sichern eines Datenblatts als SPSS-Datendatei

Wenn eine neu erstelltes Datenblatt über das Ende der Sitzung hinaus erhalten bleiben soll, muss es explizit auf einen permanenten Datenträger gesichert werden, wobei eine **SPSS-Datendatei** entsteht. In späteren Sitzungen kann durch *Öffnen* dieser Datendatei der gesicherte Zustand des Datenblatts wiederhergestellt werden.

Zwar enthält Ihre aktuelle Arbeitsdatei (das aktive und vermutlich einzige Datenblatt) noch keine Daten, aber im Deklarationsteil befinden sich bereits wertvolle Informationen, deren Verlust recht schmerzlich wäre. Daher sollten Sie schon jetzt die temporäre Arbeitsdatei in eine permanente SPSS-Datendatei sichern, indem Sie den folgenden Menübefehl wählen:

Datei > Speichern unter...

In der erscheinenden Dialogbox ist für die zu erzeugende SPSS-Datendatei ein Laufwerk, ein Verzeichnis und ein Name anzugeben:

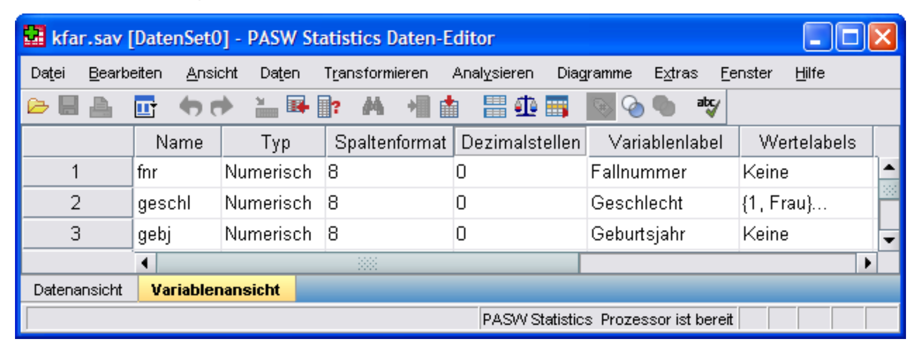


Wenn Sie die für SPSS-Datendateien vorgegebene Namenserweiterung **.sav** beibehalten, geht das spätere Öffnen besonders bequem.

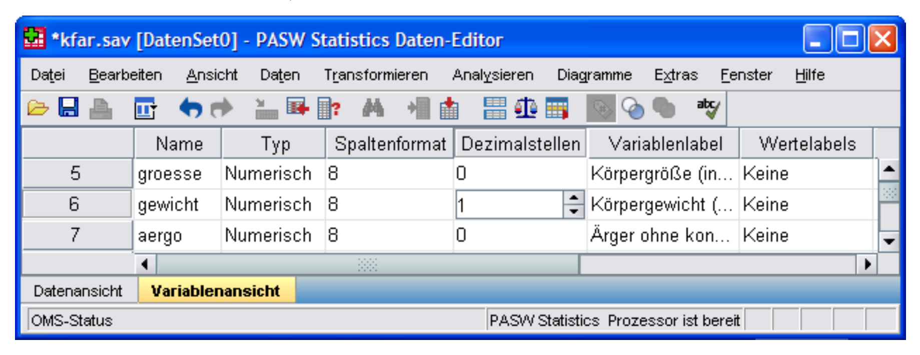
Als Name für unsere Beispieldatei wird **kfar.sav** vorgeschlagen, verbunden mit der Versicherung, die Begründung für das **r** im nächsten Abschnitt nachzuliefern.

Wenn Sie an einem Pool-PC an der Universität Trier arbeiten, können Sie die Datei im Ordner **U:\Eigene Dateien\SPSS** speichern, der beim ersten SPSS-Einsatz automatisch angelegt wurde.

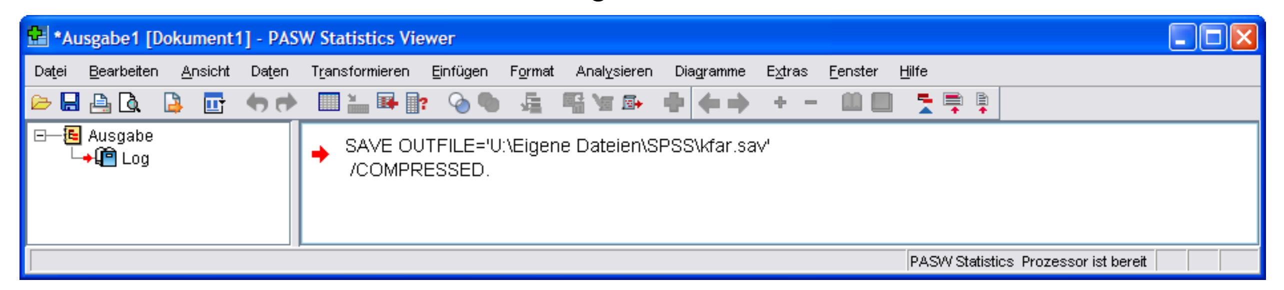
Nach dem **Speichern** zeigt die Titelzeile des Datenfensters neben dem Datenblattnamen auch den Namen der nunmehr zugeordneten Datendatei, in unserem Fall also **kfar.sav**:



Sobald ein Datenblatt gegenüber dem zuletzt gespeicherten Zustand geändert wurde, erscheint ein Sternchen vor dem Dateinamen, z.B.:



Beim Speichern führt der SPSS-Prozessor das Kommando SAVE aus, was später noch zu erläutern ist (siehe Abschnitt 6.7.1). Weil SPSS per Voreinstellung ausgeführte Kommandos protokolliert, erscheint überraschend früh ein Ausgabefenster:

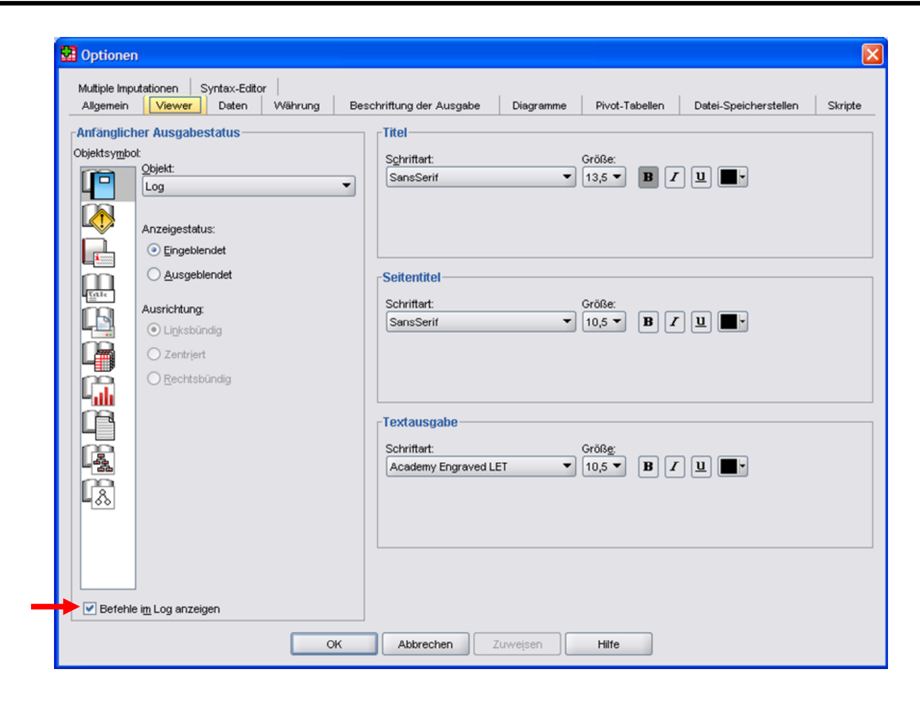


Nähere Informationen zu den Ausgabefenstern folgen in den Abschnitt 4.4 und 7.7.

Nach dem Menübefehl

Bearbeiten > Optionen

kann man im **Optionen**-Dialog auf der Registerkarte **Viewer** das Protokollieren der Kommandos abschalten:



Beim Speichern eines Datenblatts können auch alternative Dateiformate gewählt werden (z.B. EXCEL, SAS, Stata, Text).

Zum späteren Sichern in eine bereits zugeordnete Datei dient der Befehl:

Datei > Speichern

Alternativ können Sie mit der Maus auf das Symbol klicken oder die Tastenkombination **Strg+S** benutzen.

3.2.6 Rohdatendatei, Transformationsprogramm und Fertigdatendatei

Möglicherweise haben Sie sich beim Lesen des letzten Abschnitts gefragt, was das **r** im vorgeschlagenen Dateinamen **kfar.sav** bedeuten soll. Bei der Beantwortung dieser Frage sind leider einige Vorgriffe auf spätere Abschnitte nötig. Versuchen wir es trotzdem. Das **r** soll signalisieren, dass in dieser Datei die nach den Vorschriften des Kodierplans erfassten **R**ohdaten stehen. In **kfar.sav** sollen also ausschließlich folgende Arbeitsschritte einfließen:

- Variablendeklaration gemäß Kodierplan
- Datenerfassung gemäß Kodierplan
- Nötigenfalls spätere Korrekturen von Erfassungsfehlern

Damit ist diese Datei für viele im Demoprojekt geplante Auswertungsschritte noch nicht geeignet. Es fehlt z.B. der Optimismus-Testwert, welcher aus den zwölf LOT-Fragen berechnet werden muss.

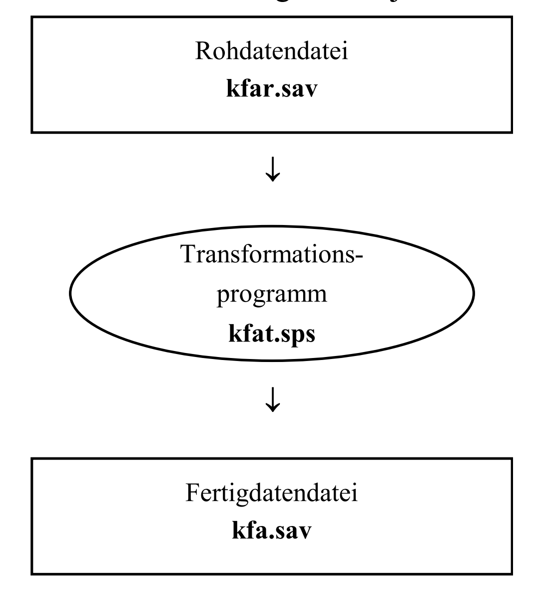
Aus der Rohdatendatei werden wir bald eine **Fertigdatendatei** herstellen, in die alle projektweit relevanten Variablenmodifikationen und -neuberechnungen einfließen sollen, so dass sie eine bequeme Datenbasis für alle statistischen und graphischen Analysen darstellt. In fast jedem Projekt sind Variablenmodifikationen und -neuberechnungen in erheblichem Umfang erforderlich.

Profis modellieren dabei nicht „per Hand“ so lange an der Rohdatei herum, bis die Fertigdatei entstanden ist, sondern sie erstellen, z.B. durch Konservieren von bearbeiteten Dialogboxen, ein so genanntes **SPSS-Programm** (siehe unten), das alle Transformationen erledigt und das bei Bedarf auch wiederholt ausgeführt werden kann.

Die zweistufige Projektdatenverwaltung mit Roh- und Fertigdatei verhindert in Kombination mit dem vermittelnden SPSS-Transformationsprogramm, dass bei jeder Änderung der Rohdaten die erwähnten Transformationen zur Fertigdatei wiederholt werden müssen. Solche Änderungen der Rohdaten (z.B. durch Fehlerkorrekturen oder Stichprobenerweiterungen) sind eher die Regel als die Ausnahme.

Weil die Kommandos des Transformationsprogramms auch mit Hilfe von korrespondierenden Dialogboxen erstellt werden können, erfordert die professionelle Vorgehensweise kaum Programmierkenntnisse.

Es wird also folgende Struktur für die Verwaltung der Projektdaten vorgeschlagen:



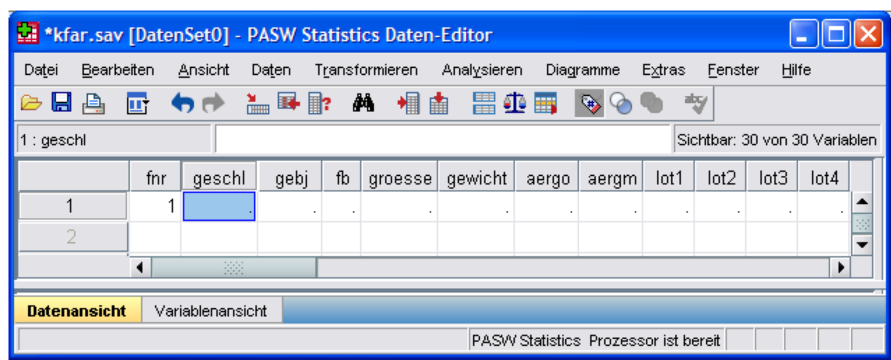
Die Erläuterungen in diesem Abschnitt werden vermutlich erst dann voll verständlich, wenn Sie sich mit Variablentransformationen und SPSS-Programmen auskennen.

Nach diesem Vorausblick wenden wir uns wieder der aktuellen Aufgabe zu: Wir tragen die erhobenen Daten in die Rohdatendatei **kfar.sav** ein.

3.2.7 Dateneingabe

Wechseln Sie bei Bedarf zur **Datenansicht** des (vermutlich noch einzigen) Dateneditorfensters, und geben Sie die Daten des ersten Falles ein:

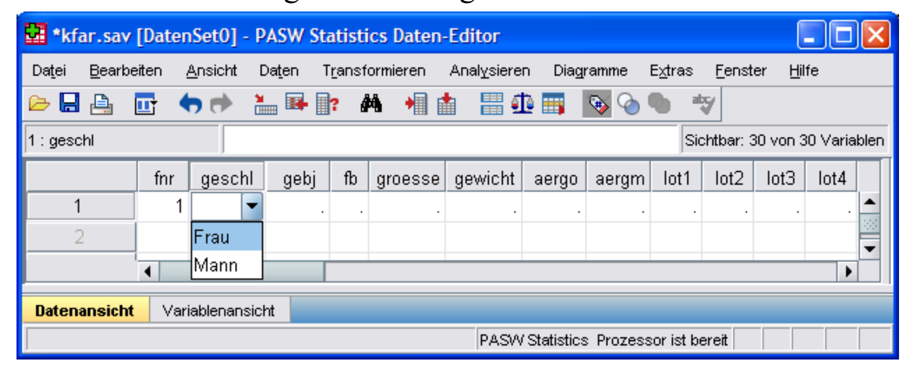
- Aktivieren Sie nötigenfalls die Zelle zur ersten Variablen, und tippen Sie den zugehörigen Wert ein.
- Drücken Sie die Taste mit dem Rechtspfeil → oder die **Tabulator**-Taste ⇆, um den eingetippten Wert zu quittieren und die Zellenmarkierung um eine Spalte nach *rechts* zu verschieben (zur nächsten Variablen):



Auch die **Enter**-Taste quittiert den eingetippten Wert, bewegt jedoch anschließend die Zellenmarkierung um eine Zeile nach *unten* (zum nächsten Fall), was in unserer jetzigen Lage weniger praktisch ist.

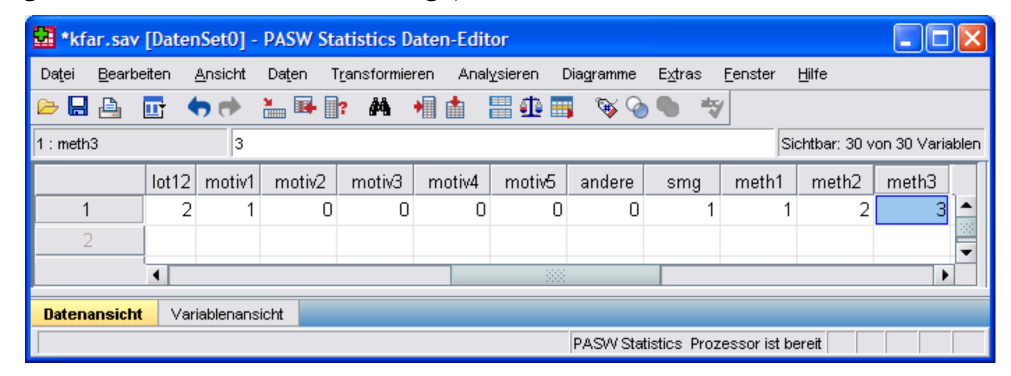
Wenn Sie auf Abwege geraten sind, können Sie die Zellenmarkierung jederzeit per Mausklick neu positionieren.

- Sobald für einen neuen Fall die erste Variablenausprägung eingetragen und quittiert wurde, erhält er für die restlichen Variablen den Initialisierungswert SYSMIS (dargestellt durch einen Punkt).
- Wenn über den Menübefehl **Ansicht > Wertelabels** die Anzeige von Wertelabels aktiviert worden ist, lässt sich z.B. nach einem Doppelklick auf die GESCHL-Zelle ein Drop-Down - Menü zur Unterstützung der Werteingabe öffnen:



So ist eine Datenerfassung ohne Kenntnis der Kodiervorschriften möglich, wobei allerdings der Zeitaufwand steigt.

- Tragen Sie die restlichen Werte des ersten Falles ein, jeweils quittiert mit der Tabulatortaste. So sieht der vollständig erfasste erste Fall unserer Stichprobe im Datenfenster aus (bei abgeschalteter Wertelabels-Anzeige):



- Wenn Sie den Wert der letzten Variablen mit der Tabulatortaste quittieren, setzt SPSS freundlicherweise die Zellenmarkierung gleich in die erste Datenzelle des nächsten Falles, so dass Sie die Dateneingabe unmittelbar fortsetzen können.

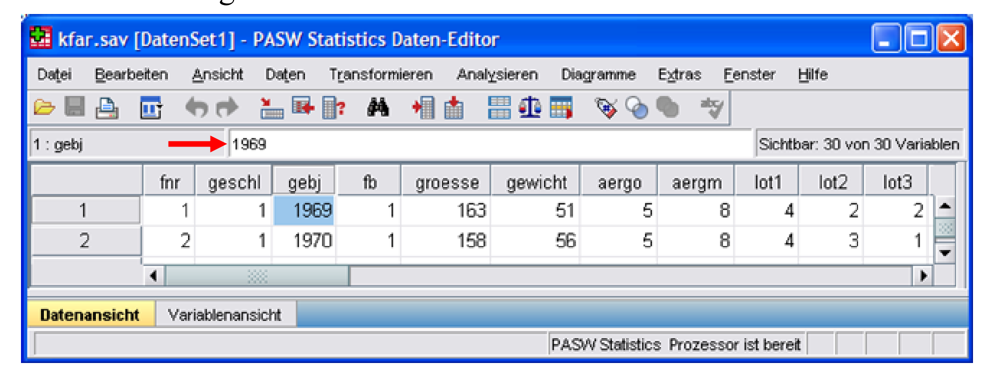
3.2.8 Daten korrigieren

3.2.8.1 Wert einer Zelle ändern

Natürlich können die Eintragungen in einer Zelle jederzeit korrigiert werden:

- Wert ersetzen:
  - Zelle markieren
  - neuen Wert eintippen, wobei der alte überschrieben wird

- Wert in der Eingabezone editieren:
  - o Zelle markieren
  - o Wert in der Eingabezone oberhalb der Datenmatrix editieren



geschehen kann

- Wert in der Zelle editieren:
  - o Doppelklick auf die Zelle
  - o Vermutlich ist ein Programmfehler dafür verantwortlich, dass die Zelle nochmals angeklickt werden muss.
  - o Nun kann der Wert in der Zelle geändert werden.

3.2.8.2 Einen Fall einfügen

Gehen Sie folgendermaßen vor, um einen Fall einzufügen:

- Setzen Sie einen rechten Mausklick auf die (von SPSS gesetzte) Zeilennummer desjenigen Falles an, *vor* dem ein neuer Fall eingefügt werden soll. Daraufhin wird die gesamte angeklickte Zeile markiert, und es erscheint ein Kontextmenü.
- Wählen Sie aus dem Kontextmenü die Option **Fälle einfügen**

Der neue Fall erhält bei allen Variablen den Initialisierungswert SYSMIS.

3.2.8.3 Einen Fall löschen

Gehen Sie folgendermaßen vor, um einen Fall, d.h. eine Zeile der Datenmatrix, komplett zu löschen:

- Setzen Sie einen rechten Mausklick die die (von SPSS gesetzte) Zeilennummer des überflüssigen Falles. Daraufhin wird die gesamte angeklickte Zeile markiert, und es erscheint ein Kontextmenü.
- Wählen Sie aus dem Kontextmenü die Option **Löschen**

3.2.8.4 Fälle verschieben

Gehen Sie folgendermaßen vor, um Fälle per Drag & Drop (Ziehen und Ablegen) zu verschieben:

- Markieren Sie die zu verschiebenden Fälle auf Windows-übliche Weise über Mausaktionen im Nummerierungsbereich, ggf. ergänzt durch die **Strg**- oder Umschalt-Taste. Lassen Sie anschließend die Maustaste wieder los.
- Klicken Sie in der Nummerierungszone erneut auf die zu verschiebende Fallauswahl, und halten Sie dabei die linke Maustaste gedrückt.

- Bewegen Sie bei gedrückter Maustaste den Mauszeiger zum Ziel der Verschiebungsaktion. Der augenblicklich eingestellte Zielort wird von SPSS durch eine rote Linie gekennzeichnet.
- Wenn Sie die Maustaste loslassen, erscheinen die Fälle am neuen Ort.

3.2.9 Weitere Möglichkeiten des Dateneditors

Über die beschriebenen Methoden hinaus bietet der Dateneditor u.a. die Möglichkeit, beliebige rechteckige Segmente einer Datenmatrix auszuschneiden, zu kopieren und einzufügen (auch zwischen verschiedenen Datensets).

Wer derartige, relativ fehleranfällige Umordnungsmaßnahmen vornimmt, wird gelegentlich von der Möglichkeit profitieren, mit der Tastenkombination **Strg+Z**, über den Symbolschalter oder mit dem Menübefehl:

Bearbeiten > Rückgängig

die letzte Änderung rückgängig machen zu können.

In Abschnitt 4.6 wird beschrieben, wie man im Datenfenster nach Variablenausprägungen suchen kann.

3.2.10 Übung

Für die Teilnehmer(innen) des realen SPSS-Kurses steht nun die Erfassung der erhobenen Daten an. Geben Sie alle Fälle ein, und sichern Sie (auch zwischendurch) in die zugeordnete Datendatei, z.B. **U:\Eigene Dateien\SPSS\kfar.sav**.

Wer dem Vorschlag aus Abschnitt 1.4.2.4 folgend zur Erfassung der Antworten auf die offene Frage im Fragebogenteil 4b ein dynamisches und sparsames Set aus kategorialen Variablen vorgesehen hat (z.B. METH1 bis METH3), der muss nicht nur mechanisch Daten eintippen, sondern auch gelegentlich mit Kreativität und Ordnungssinn neue Methodenkategorien definieren und dokumentieren. Beim Erfassen der Daten, die in diesem Manuskript analysiert werden, entstand folgende Liste:

Kategorie	Code
Faktorenanalyse	1
Regressionsanalyse	2
Korrelationsanalyse	3
Varianzanalyse	4
Strukturgleichungsanalyse	5
Clusteranalyse	6
Diskriminanzanalyse	7
Logistische Regression	8
Conjoint-Analyse	9

Diese Tabelle vervollständigt unseren Kodierplan (vgl. Abschnitt 1.4.3.5). Es bietet sich an, die Definition der Variablen METH1 bis METH3 durch entsprechende Wertelabels zu komplettieren (vgl. Abschnitt 3.2.2.3).

4 Univariate Verteilungs- und Fehleranalysen

In diesem Abschnitt werden Sie erfahren, wie schnell und bequem mit SPSS numerische und graphische Analysen durchgeführt werden können. Wir werden unsere Daten mit Hilfe deskriptiver Auswertungsmethoden sorgfältig auf Erfassungsfehler untersuchen. Dabei schlagen wir zwei Fliegen mit einer Klappe, denn eine sorgfältige Verteilungsanalyse aller Variablen gehört ohnehin zu unseren Pflichtaufgaben. Bei vielen Variablen sind wir sogar ausgesprochen neugierig auf die Verteilung (z.B. bei den Variablen AERGO und AERGM).

In manchen Projekten wird sich die Forschungsarbeit sogar auf die Beschreibung von univariaten Verteilungen beschränken. Meist sind aber auch multivariate Zusammenhangsanalysen von Interesse.

4.1 Erfassungsfehler

Speziell bei der manuellen Datenerfassung sind Fehler praktisch unvermeidbar. Manche Fehler sind als Verstöße gegen Gültigkeitsregeln relativ leicht aufzuspüren:

Beispiel: Wenn bei der Variablen GESCHL nur die Werte 1 (für Frauen) und 2 (für Männer) erlaubt sind, dann ist z.B. der Wert 3 sofort als Erfassungsfehler erkennbar.

Weit schwieriger zu entdecken sind Fehler, die keine allgemeine Gültigkeitsregel verletzen:

Beispiel: Wenn unter der oben angegebenen GESCHL-Kodierungsvorschrift für den Untersuchungsteilnehmer Kurt Müller versehentlich der Wert 1 eingegeben wurde, dann kann dieser Fehler nur durch aufwändige Handarbeit gefunden werden.

Welcher Aufwand bei der Datenprüfung erforderlich bzw. sinnvoll ist, hängt von der verwendeten Erfassungsmethode (manuell versus automatisch) und von der Stichprobengröße ab.

4.1.1 Suche nach unzulässigen Werten

Von einem Datenerfassungsprogramm mit Plausibilitätskontrolle werden unzulässige Werte zurückgewiesen und folglich von der Datendatei fern gehalten. Bei der manuellen Erfassung (z.B. mit dem SPSS-Dateneditor) findet eine derartige Eingangskontrolle nicht statt. Eine so entstandene Datei muss daher systematisch nach Daten außerhalb der zulässigen Bereiche durchsucht werden. Dies kann allerdings ohne großen Zusatzaufwand im Rahmen der aus wissenschaftlichen Gründen ohnehin empfehlenswerten univariaten Verteilungsanalyse geschehen.

4.1.2 Überprüfung von Einzelwerten

Fehler, die gegen keine Gültigkeitsregel verstoßen, lassen sich nur mit Fleißarbeit entdecken, wobei z.B. die erfassten Daten Wert für Wert mit den schriftlichen Unterlagen verglichen werden.

Eine aufwändige Prüfmethode ist *bei kleinen Stichproben* durchaus empfehlenswert, denn:

- Der Zeitaufwand ist erträglich.
- Erfassungsfehler wirken sich hier besonders stark aus.

Wir wollen exemplarisch den Effekt von Erfassungsfehlern auf die Varianz (Unsicherheit) eines Stichprobenmittelwerts als Schätzer für den zugehörigen Populationserwartungswert untersuchen. Für n erfasste Werte X_i ($i = 1, .., n$) nehmen wir an, dass sie jeweils mit einem Erfassungsfehler F_i belastet sind, wobei die Erfassungsfehler den Erwartungswert Null haben sowie untereinander und von den korrekten Beobachtungswerten T_i unabhängig sind:

$$X_i = T_i + F_i, \quad \mathrm{E}(F_i) = 0, \quad \mathrm{E}(X_i) = \mathrm{E}(T_i) = \mu$$

$$\mathrm{Var}(T_i) = \sigma^2, \quad \mathrm{Var}(F_i) = \sigma_F^2$$

Für die Varianz des Mittelwerts aus den fehlerfrei erfassten Werten gilt:

$$\mathrm{Var}(\overline{T}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} T_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(T_i) = \frac{1}{n^2}\, n\,\sigma^2 = \frac{\sigma^2}{n}$$

Für die Varianz des Mittelwerts der fehlerhaft erfassten Werte erhalten wir:

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}(T_i + F_i)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\left(\mathrm{Var}(T_i) + \mathrm{Var}(F_i)\right) = \frac{1}{n^2}\, n\left(\sigma^2 + \sigma_F^2\right) = \frac{\sigma^2}{n} + \frac{\sigma_F^2}{n}$$

Offenbar hängt der Präzisionsverlust im Stichprobenmittel, das als Schätzwert für den Erwartungswert in der Population dient, von der Erfassungsfehlervarianz σ_F^2 und von der Stichprobengröße n ab. Während sich in einer großen Stichprobe der niedrige Ausgangswert $\frac{\sigma^2}{n}$ der Unsicherheit nur unwesentlich erhöht, kommt es in einer kleinen Stichprobe mit ihrem bereits ungünstigen Ausgangsniveau zu einem erheblichen Präzisionsverlust. Als unerwünschte Folgen stellen sich ein:

- Unpräzise Parameterschätzungen
- Reduzierte Power bei Hypothesentests

Obwohl bei unserer kleinen Stichprobe eine Einzelprüfung aller Werte angemessen wäre, verzichten wir aus Zeitgründen darauf. Es gehört übrigens zu den lehrreichen Erfahrungen der realen SPSS-Kurse, dass die selbständig als Untersuchungsleiter agierenden Teilnehmer aus Kopien desselben Fragebogenstapels aufgrund individueller Erfassungsfehler recht unterschiedliche Ergebnisse ermitteln (auch bei den zentralen Hypothesentests).

4.2 Öffnen einer SPSS-Datendatei

Vermutlich haben Sie nach der anstrengenden Datenerfassung eine Pause eingelegt und SPSS verlassen, so dass wir jetzt offiziell die Fortsetzung einer unterbrochenen Projektarbeit üben können. Starten Sie SPSS, und öffnen Sie Ihre vorhandene Rohdatendatei **kfar.sav**, entweder mit Hilfe des Startassistenten oder über den Menübefehl

Datei > Zuletzt verwendete Daten

Bei einer für Mausaktionen zugänglichen Datendatei stehen weitere Techniken zum Öffnen zur Verfügung:

- Doppelklick
- Drag & Drop: Ziehen und auf einem Dateneditorfenster ablegen

Beim Öffnen einer Datendatei legt SPSS ein neues Datenblatt an und kopiert die eingelesenen Daten samt Variablendeklarationen dorthin. Alle Veränderungen, die Sie in der Datenmatrix oder im Deklarationsteil vornehmen, wirken sich zunächst nur auf das temporäre Datenblatt aus. Gegebenenfalls müssen Sie also diese Änderungen über den Menübefehl

Datei > Speichern

in die permanente SPSS-Datendatei **kfar.sav** übernehmen.

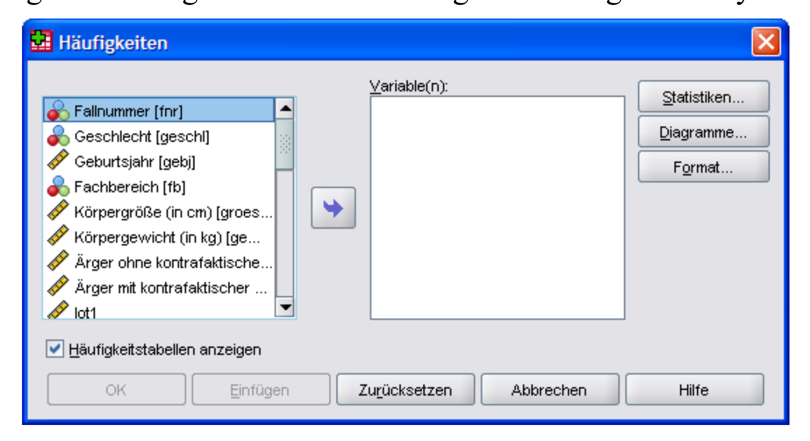
4.3 Verteilungsanalysen anfordern

Da wir unsere Daten mit dem SPSS-Dateneditor erfasst haben, der keine Plausibilitätskontrolle bei der Eingabe vornimmt, müssen wir nach den Überlegungen aus Abschnitt 4.1 systematisch nach unzulässigen Werten suchen. Die meisten der dazu erforderlichen deskriptiven Datenanalysen wären im Rahmen der routinemäßigen Verteilungsuntersuchungen ohnehin erforderlich.

Wir untersuchen zunächst die Verteilungen der nominalskalierten Variablen GESCHL und FB mit Hilfe von Häufigkeitstabellen und Balkendiagramme. Mit dem Menübefehl

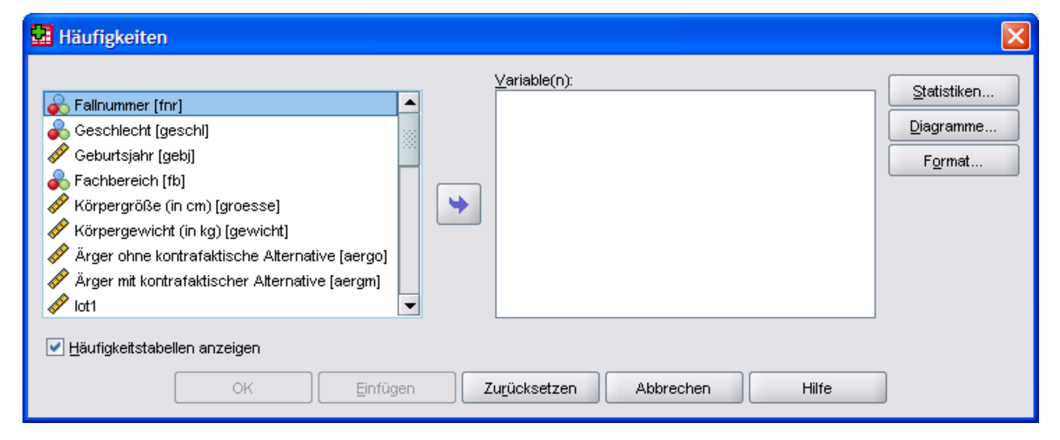
Analysieren > Deskriptive Statistik > Häufigkeiten...

erhalten wir die folgende Dialogbox zur Anforderung von Häufigkeitsanalysen:

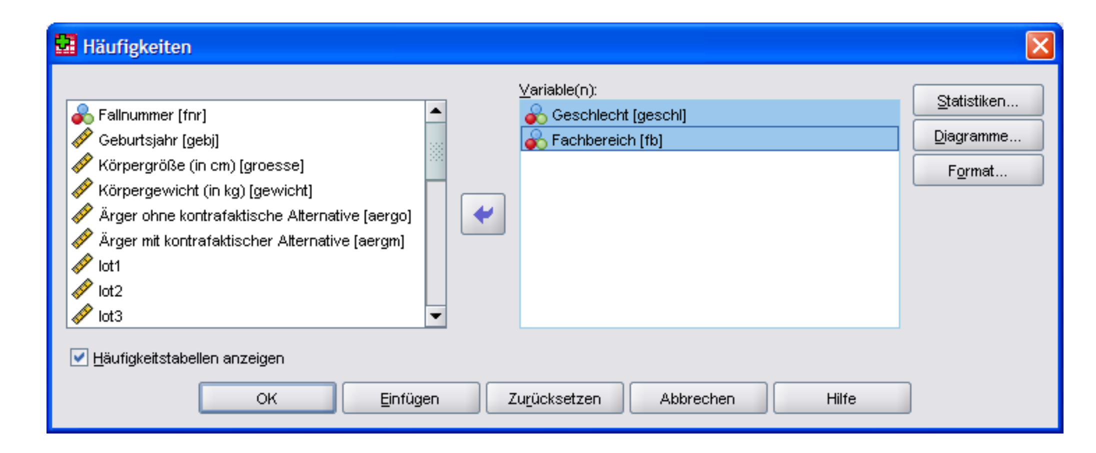


Zur bequemen Spezifikation der im aktuellen Prozeduraufruf zu analysierenden Variablen dienen die beiden Variablen-Auswahlbereiche. Links stehen alle Variablen der Arbeitsdatei, die derzeit *nicht* für die Analyse ausgewählt sind (*Anwärterliste*). Rechts daneben, im Bereich **Variable(n)**, stehen die Ausgewählten (*Teilnehmerliste*). Dazwischen befindet sich ein Transportschalter, mit dem sich links markierte Variablen nach rechts und rechts markierte Variablen nach links verschieben lassen.

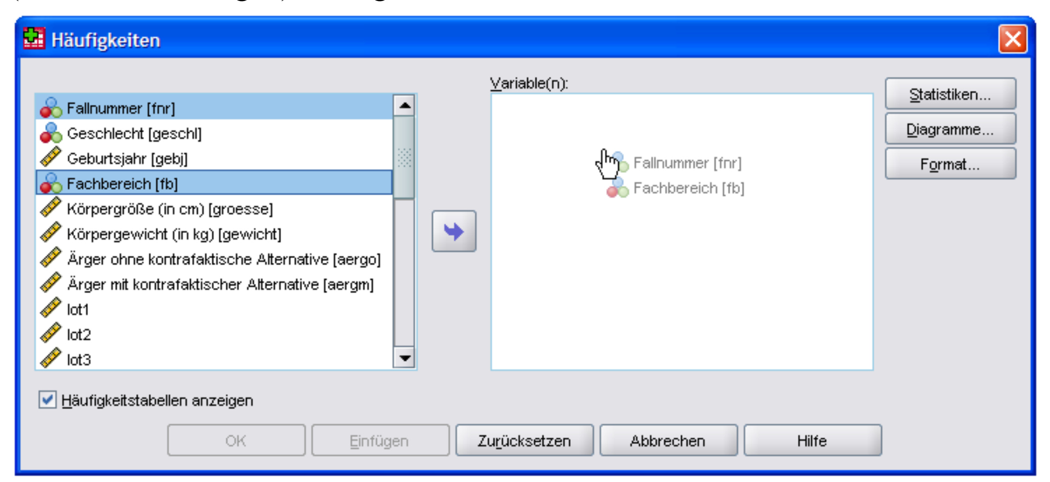
SPSS-Dialogboxen lassen sich vergrößern, so dass man im konkreten Beispiel für die komplette Anzeige der Variablenlabels sorgen kann:



Markieren Sie in der Anwärterliste (links) die Variablen GESCHL und FB, und befördern Sie diese per Mausklick auf den Transportschalter in die Teilnehmerliste (rechts):

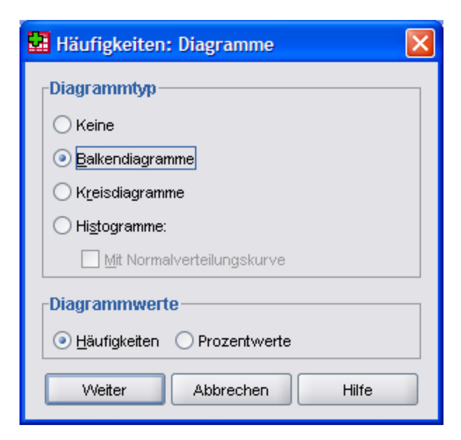


Statt den Schalter zu benutzen, können Sie in SPSS solche Transportaufgaben auch per Drag & Drop (Ziehen und Ablegen) erledigen:



Bei einer längeren Liste ist es sehr hilfreich, dass SPSS beim Eintippen einer Variablenbezeichnung das erste aktuell kompatible Listenelement markiert, wobei der Name oder das Label (falls vorhanden) anzugeben ist.

Begeben Sie sich anschließend in die Subdialogbox **Diagramme**, und wählen Sie im Rahmen **Diagrammtyp** die Option **Balkendiagramme**, weil die Merkmale Geschlecht und Fachbereich nominalskaliert sind:



Wer nicht mehr genau weiß, wozu man Balkendiagramme und Histogramme verwendet, kann sich mit der kontextsensitiven **Hilfe** Aufklärung verschaffen.

Quittieren Sie die Subdialogbox **Diagramme** mit **Weiter** und die Hauptdialogbox mit **OK**. Daraufhin präsentiert SPSS die Ergebnisse im Ausgabefenster (**PASW Statistics Viewer**), das sich in den Vordergrund drängt.

Wir erfahren in der ersten Tabelle, dass bei den untersuchten Variablen alle Werte vorhanden waren:

Statistiken

		Geschlecht	Fachbereich
N	Gültig	31	31
	Fehlend	0	0

Bei Anforderung einer Häufigkeitsanalyse produziert SPSS per Voreinstellung eine Tabelle, die für jeden aufgetretenen Wert eine Zeile mit folgenden Angaben enthält:

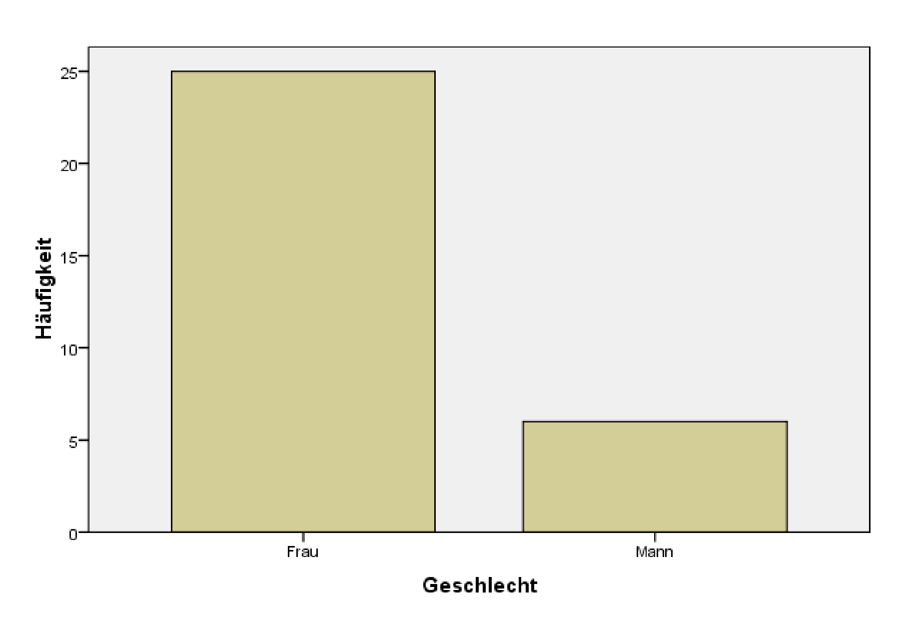
- Absolute Häufigkeit
- Prozentualer Anteil am Stichprobenumfang
- Prozentualer Anteil an den validen Werten (ohne MD-deklarierte Werte)
- kumulativer valider Prozentanteil (Anteil valider Werte, die nicht größer sind)

Für GESCHL erhalten wir diese Häufigkeitstabelle:

Geschlecht

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Frau	25	80,6	80,6	80,6
	Mann	6	19,4	19,4	100,0
	Gesamt	31	100,0	100,0	

und das folgende Balkendiagramm:



Zunächst beobachten wir, dass bei der Variablen GESCHL kein unzulässiger Wert vorliegt.

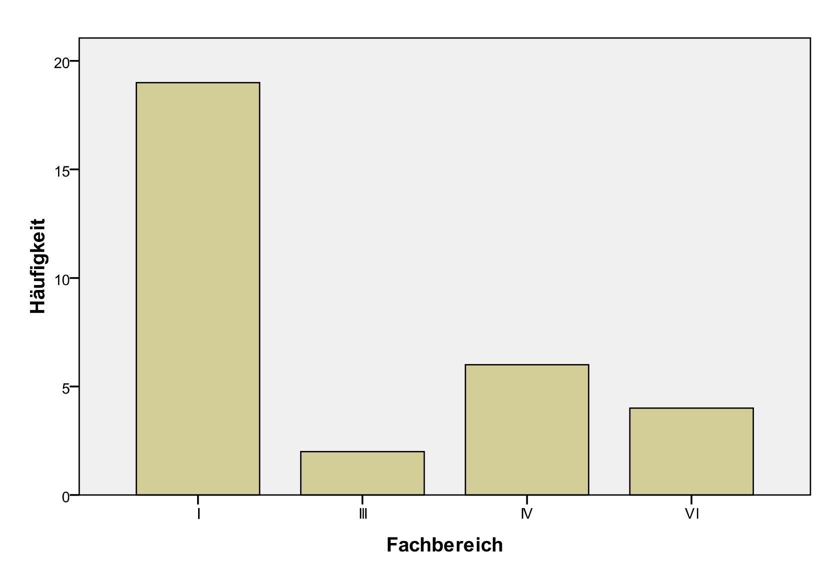
Bei der Geschlechtsverteilung stellen wir einen sehr hohen Frauenanteil fest, der als wesentliches Merkmal unserer Stichprobe berichtet werden muss. Bei potentiell geschlechtsabhängigen Ergebnissen müssen wir besonders vorsichtig interpretieren und generalisieren.

Erste Hinweise zur Ursache der hohen Frauenquote liefert die empirische Verteilung der Fachbereichs-Variablen:

Fachbereich

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	I	19	61,3	61,3	61,3
	III	2	6,5	6,5	67,7
	IV	6	19,4	19,4	87,1
	VI	4	12,9	12,9	100,0
	Gesamt	31	100,0	100,0	

Fachbereich



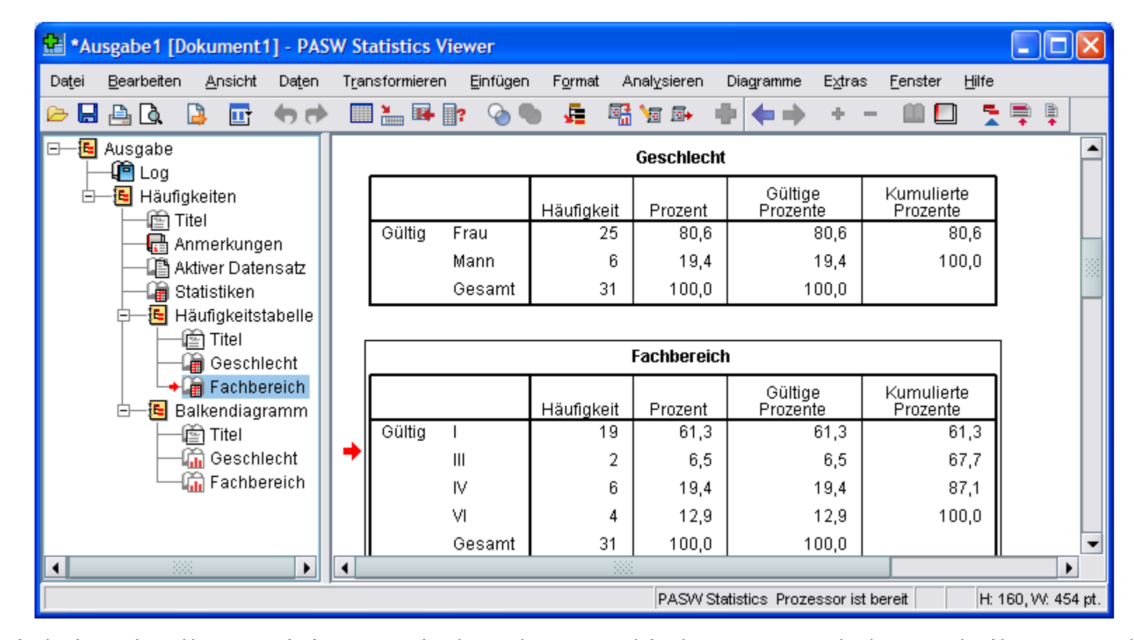
Wir sehen, dass im SPSS-Kurs, der die Manuskriptdaten geliefert hat, der Fachbereich I sehr stark vertreten war, was mit dem Kurstermin zusammenhängen mag. Im Fachbereich I der Universität Trier (Fächer: Philosophie, Pädagogik, Psychologie) ist der Frauenanteil sehr hoch.

Der aktuelle Abschnitt sollte nur einen ersten Eindruck von den Graphikmöglichkeiten des SPSS-Systems vermittelt. Wir haben eine integrierte Graphik-Option der Dialogbox zur Häufigkeitsanalyse benutzt. Die meisten graphischen Darstellungsmöglichkeiten bietet SPSS über das Hauptmenü **Diagramme** an, mit dem wir uns später befassen werden.

Die obigen SPSS-Ausgaben wurden übrigens über die Windows-Zwischenablage in das Textverarbeitungsprogramm Microsoft Word© übertragen. Mit dieser Form des Datenaustauschs und mit anderen Optionen beim Arbeiten mit dem Ausgabefenster beschäftigen wir uns im nächsten Abschnitt.

4.4 Arbeiten mit dem Ausgabefenster (Teil I)

In seiner voreingestellten Variante ist das SPSS-Ausgabefenster, das auch als **Viewer** bezeichnet wird, zweigeteilt in den Navigationsbereich (die Gliederungsansicht) am linken Rand und den eigentlichen Inhaltsbereich:



So wird ein schnelles Navigieren zwischen den verschiedenen Ausgabebestandteilen ermöglicht.

Die Aufteilung des verfügbaren Platzes auf die beiden Teile des Viewers kann per Maus beliebig verändert werden: Trennlinie anklicken und bei gedrückter Maustaste verschieben.

Wesentliche Bestandteile des Inhaltsbereichs sind Pivot-Tabellen und Graphiken. Zu ihrer Nachbearbeitung steht jeweils ein spezieller Editor zur Verfügung, der per Doppelklick auf ein Objekt gestartet wird (siehe unten). Außerdem können in einem Viewer-Dokument noch protokollierte SPSS-Kommandos, Textausgaben, Warnungen, Anmerkungen und Titelzeilen auftreten.

4.4.1 Arbeiten im Navigationsbereich

Die meisten der anschließend beschriebenen Aktionen im Navigationsbereich wirken sich synchron auch auf den Inhaltsbereich aus.

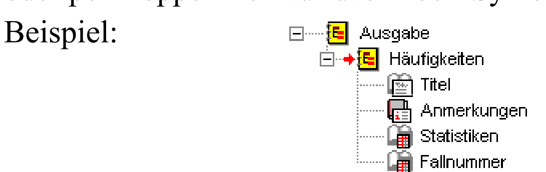
4.4.1.1 Fokus positionieren

Ein kleiner roter Pfeil zeigt im Gliederungsbereich auf die Bezeichnung derjenigen Ausgabe, die gerade im Inhaltsbereich privilegiert dargestellt wird. Per Mausklick auf eine andere Ausgabenbeschriftung kann dieser Fokus verschoben werden.

4.4.1.2 Ausgabeblöcke bzw. Teilausgaben aus- oder einblenden

Ein *Block* mit zusammengehörigen Ausgaben (in der Regel entstanden aus einer Analyseanforderung) wird ...

- ausgeblendet: per Mausklick auf das Minus-Zeichen neben dem Block-Symbol oder per Doppelklick auf das Block-Symbol.
  Beispiel:



- eingeblendet: per Mausklick auf das Plus-Zeichen neben Block-Symbol oder per Doppelklick auf das Block-Symbol.
  Beispiel:



Eine *Teil*ausgabe innerhalb eines Blocks wird per Doppelklick auf das zugehörige Buchsymbol aus- bzw. eingeblendet. Das Buchsymbol erscheint dementsprechend zugeklappt (im Beispiel: Anmerkungen) oder aufgeklappt (im Beispiel: Statistiken).

4.4.1.3 Ausgabeblöcke oder -teile markieren

Im Navigationsbereich können Sie auf Windows-übliche Weise Ausgabeblöcke und/oder Teilausgaben markieren:

- Einen Ausgabeblock: Per Mausklick auf das Block-Symbol oder auf die Beschriftung
- Eine Teilausgabe: Per Mausklick auf das Buchsymbol oder auf die Beschriftung
- Mehrere Blöcke bzw. Teile: Durch Mausklicks in Kombination mit der ⇧-bzw. **Strg**-Taste

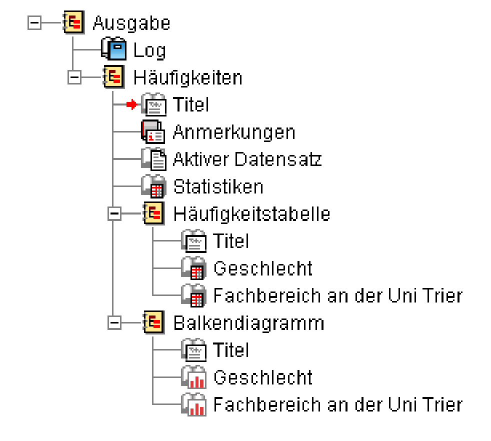
4.4.1.4 Blöcke bzw. Teilausgaben kopieren, verschieben oder löschen

Sie können markierte Blöcke bzw. Teilausgaben ...

- Löschen: mit der **Entf**-Taste oder dem Menübefehl **Bearbeiten > Löschen**
- Kopieren bzw. Verschieben:
  - mit der Maus: Ziehen und Ablegen, beim *Kopieren* zusätzlich *vor* Beginn der Bewegung die **Strg**-Taste drücken
  - via Zwischenablage: mit den Items aus dem Menü **Bearbeiten** oder den äquivalenten Tastenkombinationen: **Kopieren** bzw. **Ausschneiden**, Ziel markieren und **Einfügen**

4.4.1.5 Befördern und Degradieren

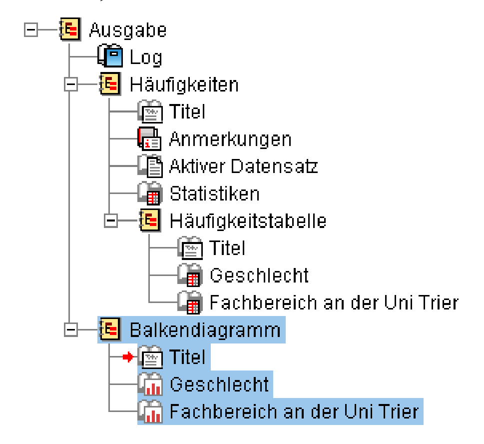
Die Ausgabeblöcke in einem Viewer-Dokument müssen sich nicht unbedingt auf derselben Gliederungsebene befinden, sondern können baumartig angeordnet werden. Von dieser Strukturierungsmöglichkeit macht z.B. die Prozedur zur Häufigkeitsanalyse Gebrauch:



Ausgabeblöcke können mit den Pfeiltasten in der Symbolleiste **Gliederung**



„befördert“ oder „degradiert“ werden, z.B.:

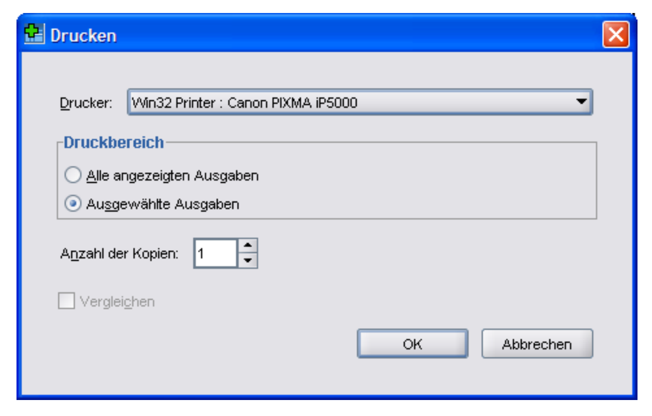


4.4.2 Viewer-Dokumente drucken

Über den Menübefehl

Datei > Drucken

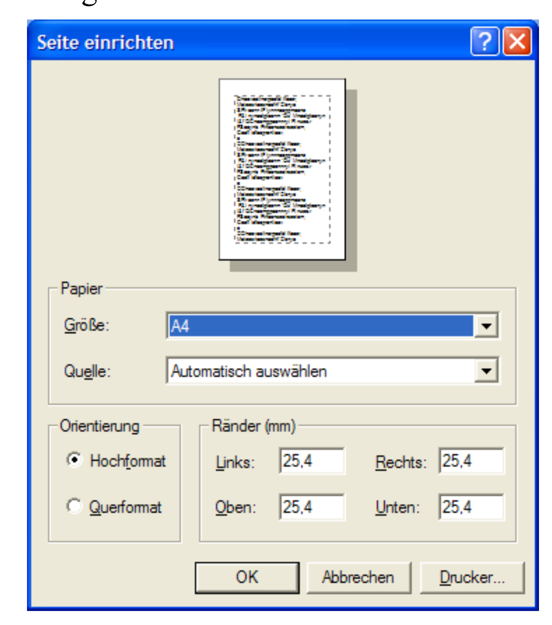
können Sie alle angezeigten oder alle markierten Ausgabebestandteile drucken, z.B.:



Nach dem Menübefehl

Datei > Seite einrichten

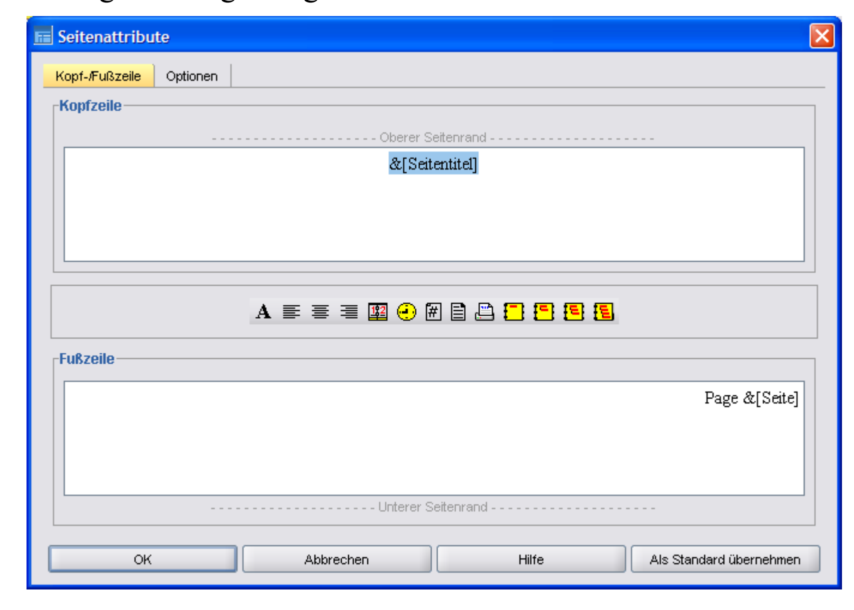
kann man in der folgenden Dialogbox das Seitenformat beeinflussen:



Zur Gestaltung der Ausgabe finden sich nach

Datei > Seitenattribute

in der folgenden Dialogbox einige Möglichkeiten:



Den Erfolg Ihrer Bemühungen können Sie über **Datei > Seitenansicht** auch schon vor dem Ausdruck begutachten.

4.4.3 Ausgaben sichern und öffnen

Zum Speichern eines Viewer-Dokuments dienen die Menübefehle **Datei > Speichern unter** bzw. **Datei > Speichern**. Dabei entstehen Viewer-Dateien, die üblicherweise durch die Namenserweiterung **.spv** gekennzeichnet werden. SPSS-Ausgaben sollten z.B. *dann* in elektronischer Form gespeichert werden, wenn sie (auszugsweise) in Dokumente anderer Programme eingegangen sind, z.B. in MS-Word - Dateien. Mit SPSS ist eine nachträgliche Modifikation dieser Ausgaben leicht möglich, mit den Fremdprogrammen aber kaum.

Zum Öffnen eines Viewer-Dokuments mit den Befehlen **Datei > Öffnen > Ausgabe** oder **Datei > Zuletzt geöffnete Dateien** gibt es nichts Ungewöhnliches zu berichten.

4.4.4 Objekte via Zwischenablage in andere Anwendungen übertragen

Mit der Tastenkombination **Strg+C** oder mit dem Menübefehl

Bearbeiten > Kopieren

fordert man SPSS auf, die markierten Ausgabeobjekte (z.B. Tabellen und/oder Diagramme) in die Windows-Zwischenablage zu befördern.

Zum Einfügen in der Zielanwendung kann man den Menübefehl

Bearbeiten > Einfügen

bzw. die Tastenkombination **Strg+V** verwenden.

SPSS legt die Daten in mehreren Formaten in der Zwischenablage ab, und je nach Zielanwendung kann es sinnvoll sein, über den Menübefehl

Bearbeiten > Inhalte Einfügen

auf das entnommene Format Einfluss zu nehmen. Wenn Sie beim Einfügen einer Tabelle das Format **Grafik (Windows-Metadatei)** oder das Format **Bild (Erweiterte Metadatei)** wählen, erhalten Sie in der Zielanwendung ein Graphik-Implantat mit dem Original-Design aus dem SPSS-Viewer. So wurden z.B. die in Abschnitt 4.3 wiedergegebenen Tabellen übertragen.

Bei Verwendung des voreingestellten Einfügeformats sollte z.B. aus einer SPSS-Tabelle eine MS-Word - Tabelle werden, die sich uneingeschränkt mit den Mitteln des Zielprogramms editieren lässt.

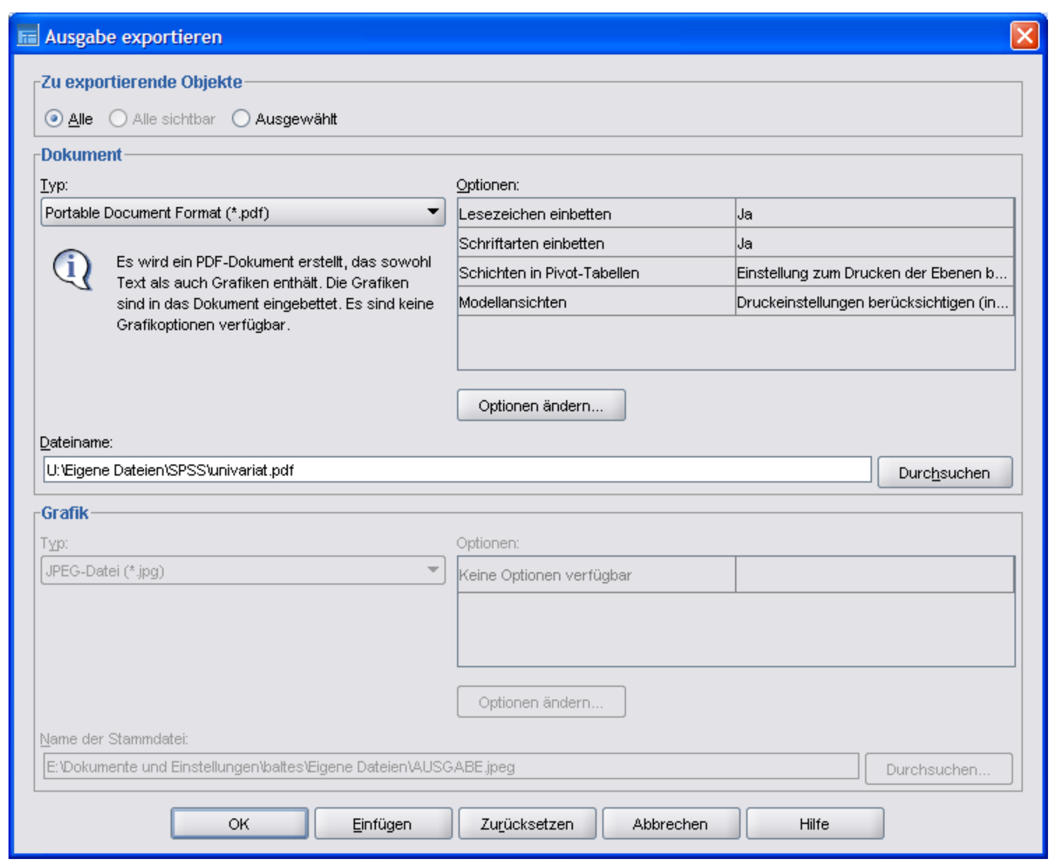
4.4.5 Ausgaben exportieren

Pivot-Tabellen, Diagramme und sonstige Ausgaben können in diversen Formaten (z.B. HTML, PDF, MS-Word/RTF, Text) exportiert werden. So lassen sich z.B. Ergebnispakete in elektronischer Form an Mitglieder einer Arbeitsgruppe übergeben, die über keine passende SPSS-Version zum Öffnen der Ausgabedateien (Namenserweiterung **spv**) verfügen.

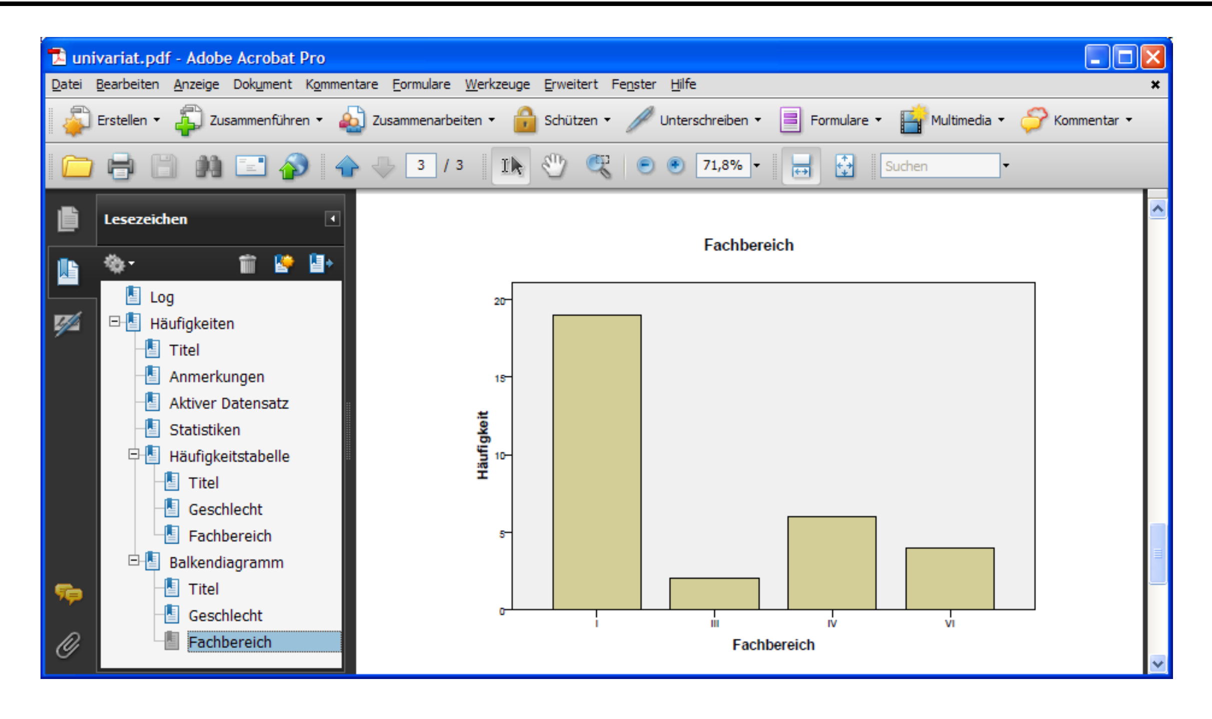
Der Export wird angefordert mit

Datei > Exportieren...

Mit folgender Dialogbox wird z.B. das gesamte Viewer-Dokument im PDF-Format exportiert:



Aus den Elementen der Viewer-Gliederungsansicht entstehen PDF-Lesezeichen:



4.4.6 Mehrere Ausgabefenster verwenden

Bislang war immer von *dem* Ausgabefenster die Rede. Im Verlauf einer längeren Auswertungsarbeit kann es der Übersichtlichkeit halber sinnvoll sein, ein zusätzliches Ausgabefenster anzufordern. Dazu dient der Menübefehl:

Datei > Neu > Ausgabe

Wenn mehrere Ausgabefenster vorhanden sind, muss geregelt werden, in welches Fenster SPSS zukünftige Ausgaben schreiben soll. Daher ist stets ein *Haupt*ausgabefenster festgelegt. Es ist an einem Pluszeichen im Symbol zum Systemmenü (siehe linken Rand der Titelzeile) sowie an einem passiven Hauptfenster-Schalter in seiner Symbolleiste zu erkennen, z.B.:



Dieser Schalter dient nämlich im aktiven Zustand dazu, ein Ausgabefenster zum *Haupt*ausgabefenster zu ernennen.

Um ein bestimmtes Ausgabefenster in den Vordergrund zu holen, können Sie es anklicken oder das **Fenster**-Menü eines beliebigen SPSS-Fensters benutzen. Jedes Ausgabefenster kann auf Windows-übliche Weise geschlossen werden, z.B. indem man es in den Vordergrund holt und dann anordnet:

Datei > Schließen

4.4.7 Übungen

1) Markieren Sie den Ausgabeblock mit der Häufigkeitsanalyse für GESCHL und FB, und löschen Sie ihn mit der **Entf**-Taste.

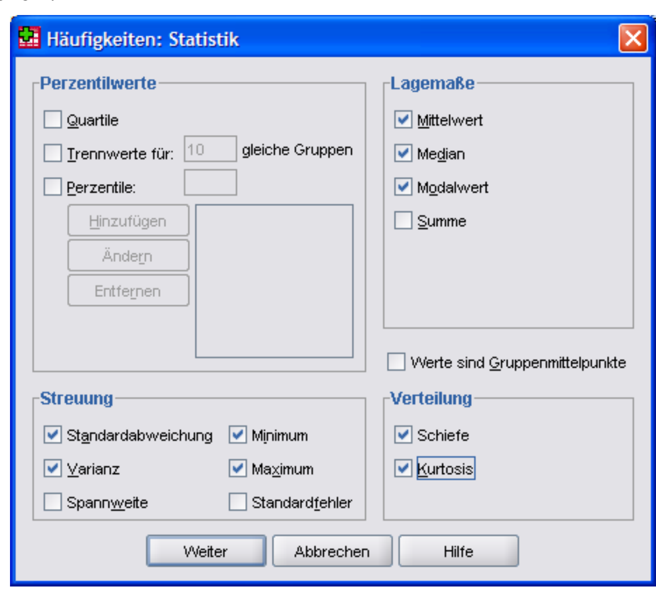
2) Öffnen Sie erneut in die Dialogbox zur Häufigkeitsanalyse. Statt den zugehörigen Menübefehl zu wiederholen, können Sie mit dem Symbol eine Liste der zuletzt benutzten Dialogboxen aufrufen und daraus per Mausklick den Eintrag **Häufigkeiten** wählen. Die Dialogbox ist noch im selben Zustand, den Sie eben verlassen haben. Dies gilt generell in SPSS, so dass Sie bei der sukzessiven Modifikation einer Anforderung innerhalb einer Sitzung jeweils auf dem letzten Stand weitermachen können.
3) Wählen Sie in der Subdialogbox **Diagramme** eine Beschriftung der Y-Achse durch **Prozentwerte**.

4.5 Häufigkeits- bzw. Fehleranalysen für die restlichen Projektvariablen

4.5.1 Übung

Führen Sie die restlichen Verteilungs- bzw. Fehleranalysen zu unserem Projekt aus. Die mehrfach benötigte **Häufigkeiten**-Dialogbox sollte jeweils über den Schalter **Zurücksetzen** von alten Einstellungen (auch in den Subdialogboxen) befreit werden.

1) Die Merkmale Geburtsjahr, Größe, Gewicht und die beiden Ärgermaße können näherungsweise als metrisch angesehen werden. Lassen Sie sich daher für die zugehörigen Variablen ausgeben:
   - Histogramme mit eingezeichneter Normalverteilungsdichte
   - keine Häufigkeitstabellen
     Das für Tabellen zuständige Kontrollkästchen in der Dialogbox **Häufigkeiten** ist per Voreinstellung markiert. Sie müssen also die Markierung durch Anklicken beseitigen.
   - folgende Statistiken: Mittelwert, Median, Modalwert, Standardabweichung, Varianz, Minimum, Maximum, Schiefe, Kurtosis (Exzeß)
     Zur Auswahl der gewünschten **Statistiken** müssen Sie die zuständige Subdialogbox per Knopfdruck aktivieren:



2) Lassen Sie sich für die LOT-Variablen ausgeben:
   - Häufigkeitstabellen
   - keine Graphiken
   - folgende Statistiken: Mittelwert, Median, Modalwert, Standardabweichung, Varianz, Minimum, Maximum

3) Lassen Sie sich für die Variablen MOTIV1 bis MOTIV5, ANDERE, SMG und METH1 bis METH3 ausgeben:
   - Häufigkeitstabellen
   - keine Graphiken
   - keine Statistiken

4) Prüfen Sie für alle Variablen nach, ob unzulässige Werte vorliegen.

5) Untersuchen Sie bei den metrischen Variablen GROESSE, GEWICHT, AERGO und AERGM zusätzlich, ob diese annähernd normal verteilt sind. Beziehen Sie in Ihr Urteil die Statistiken Schiefe und Kurtosis sowie deren Standardfehler ein.

   Die Vergleiche mit der Normalverteilung erfolgen hier aus purem Interesse an den Verteilungen der betrachteten Variablen, ohne dabei an die Verteilungs*voraussetzungen* irgendwelcher Testverfahren zu denken. Diese Voraussetzungen beziehen sich ohnehin in der Regel nicht auf die momentan von uns analysierten univariaten Verteilungen, sondern z.B. auf die Verteilungen der Residuen eines bestimmten statistischen Modells. Nähere Aussagen sind nur im Zusammenhang mit konkreten Testverfahren möglich.

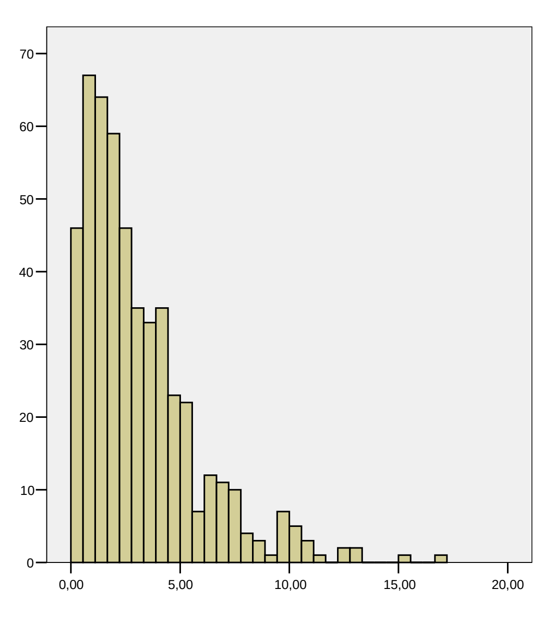
Hinweise zu den Statistiken Schiefe und Kurtosis:

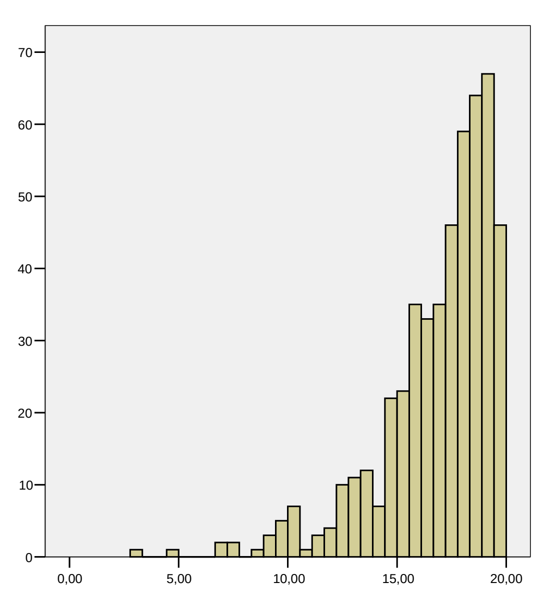
Schiefe

Bei symmetrischen Variablen hat die Schiefestatistik den Wert Null. Sie wird positiv bei linkssteil (bzw. rechtsschief) verteilten Variablen, wenn also die Verteilungsmasse am linken Rand konzentriert ist, und negativ bei rechtssteil (bzw. linksschief) verteilten Variablen, z.B.:



Negativ schiefe Verteilung





Zur Stichprobenschiefe wird auch der zugehörige Standardfehler ausgegeben, mit dessen Hilfe wir Tests zur Populationsschiefe veranstalten können. Diese sind allerdings nur approximativ gültig (bei unendlich großen Stichproben) und folglich bei kleinen Stichproben mit Vorsicht zu genießen. Ihr Vorzug gegenüber den später vorzustellenden Normalverteilungs-Anpassungstests besteht darin, dass sie gezielt auf Verletzungen der Verteilungs*symmetrie* ansprechen.

Bei einem α−Fehlerrisiko von 5 % ist die *zwei*seitige Nullhypothese, dass die Schiefe in der Population gleich Null sei, zu verwerfen, wenn gilt:

$$\frac{|\text{Schiefe}|}{\text{SF(Schiefe)}} > 1{,}96$$

Beim Wert 1,96 handelt es sich um das 97,5%-Quantil der Standardnormalverteilung.

Der Test zum gerichteten Hypothesenpaar:

$$H_0\text{: Schiefe} \geq 0 \quad \text{versus} \quad H_1\text{: Schiefe} < 0$$

entscheidet sich beim selben α–Niveau gegen seine Nullhypothese, wenn der Quotient aus der Schiefe und ihrem Standardfehler das 5%-Quantil der Standardnormalverteilung unterbietet:

$$\frac{\text{Schiefe}}{\text{SF(Schiefe)}} < -1{,}65$$

Analog lässt sich auch die einseitige Nullhypothese mit umgekehrtem Vorzeichen prüfen.[1]

Kurtosis (Exzeß)

Der Exzeß (synonym: Kurtosis, Breitgipfligkeit, Wölbung) ist bei normalverteilten Variablen gleich Null. Er wird negativ bei breiteren und positiv bei schlankeren Verteilungen. Mit Hilfe des zugehörigen Standardfehlers können analog zum Vorgehen bei der Schiefe-Statistik (siehe oben) approximativ (bei unendlich großen Stichproben) gültige Tests zum Exzeß in der Population durchgeführt werden.

4.5.2 Diskussion ausgewählter Ergebnisse

a) Die Verteilungen der zentralen KFA-Variablen (AERGO, AERGM)

Bei den zentralen KFA-Variablen (AERGO, AERGM) finden sich keine irregulären Werte:

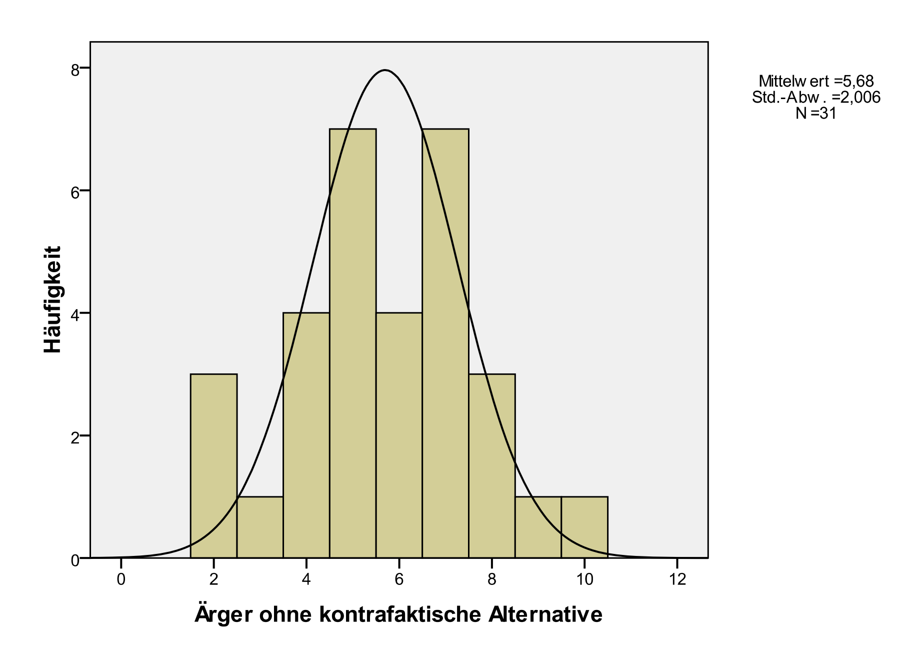
Statistiken

		Geburtsjahr	Körpergröße (in cm)	Körpergewicht (in kg)	Ärger ohne kontrafaktische Alternative	Ärger mit kontrafaktischer Alternative
N	Gültig	31	31	31	31	31
	Fehlend	0	0	0	0	0
Mittelwert		1968,94	172,81	63,48	5,68	7,68
Median		1969,00	174,00	60,00	6,00	8,00
Modus		1967[a]	176	60	5[a]	8
Standardabweichung		3,214	8,288	10,494	2,006	2,271
Varianz		10,329	68,695	110,125	4,026	5,159
Schiefe		,017	,448	1,265	-,080	-1,451
Standardfehler der Schiefe		,421	,421	,421	,421	,421
Kurtosis		,241	-,166	1,889	-,277	2,013
Standardfehler der Kurtosis		,821	,821	,821	,821	,821
Minimum		1961	158	50	2	1
Maximum		1975	192	96	10	10

a. Mehrere Modi vorhanden. Der kleinste Wert wird angezeigt.

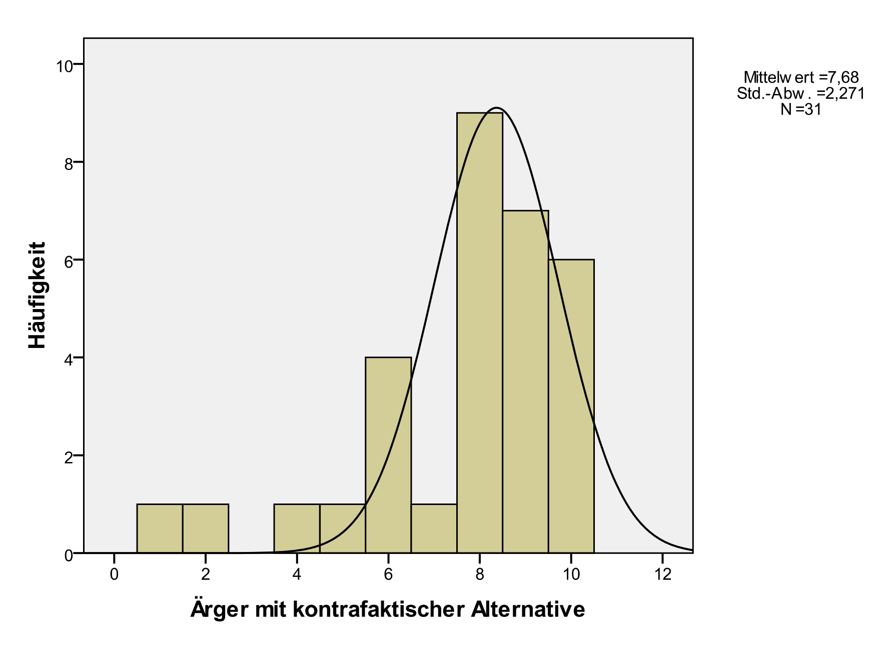
Die Verteilung der Ärgermessung in der Situation *ohne* kontrafaktische Alternative (AERGO) macht einen recht „normalen“ Eindruck:

[1] Wer in seinem Gedächtnis nicht mehr genügend Kenntnisse zur Inferenzstatistik reaktivieren konnte, der sei auf den Abschnitt 7.1 vertröstet.



Die Verteilungskennwerte Schiefe (= -0,08) und Kurtosis (= -0,277) sind nach den oben angegebenen Tests nicht signifikant von Null verschieden.

Wir sind nun sehr gespannt auf die Verteilung der Ärgermessung in der Situation *mit* kontrafaktischer Alternative (AERGM), weil sich ein KFA-Effekt hier deutlich abzeichnen sollte. Es ist generell zu empfehlen, sich mit möglichst einfachen Graphiken und Statistiken ein präzises Bild von der Effektlage zu verschaffen, statt einem Signifikanztest blind zu vertrauen, der eventuell durch technische Fehler belastet ist. Im Vergleich zur relativ symmetrischen Verteilung von AERGO um den Mittelwert 5,68 ist die AERGM-Verteilung deutlich nach rechts verschoben (Mittelwert 7,68) und „deformiert“:



Wir sehen einen mittleren Ärgeranstieg um 20° (bei Rückübersetzung in die Celsius-Skala des Fragebogens). Außerdem ist die AERGM-Verteilung am rechten Rand konzentriert und deutlich verschieden von einer Normalverteilung, was sich auch in signifikanten Ergebnissen der Tests zu Schiefe und Kurtosis widerspiegelt:

$$\frac{|\,\text{Schiefe}\,|}{\text{SF(Schiefe)}} = 3{,}447 > 1{,}96$$

$$\frac{|\text{Kurtosis}|}{\text{SF(Kurtosis)}} = 2{,}451 > 1{,}96$$

Hier sind *zwei*seitige Tests durchzuführen, weil keine gerichteten Hypothesen vorlagen. Wir haben zwar eine explizite Hypothese über die Richtung des KFA-Effekts (vgl. Abschnitt 1.3.2), doch muss die Verschiebung einer Verteilung nach rechts nicht zwangsläufig zu einer negativen Schiefe führen (siehe Abbildung in Abschnitt 1.3.2). Offenbar ist aber der KFA-Effekt so stark, dass er die Ärgerverteilung an die „Decke" geschoben und damit rechtssteil (negativ schief) gemacht hat.

b) Ergebnis der Fehleranalyse

Unsere Fehleranalyse liefert nur einen „Treffer". In der Häufigkeitstabelle zur Variablen LOT10 entdecken wir den verbotenen Wert Null:

lot10

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0	1	3,2	3,2	3,2
	1	4	12,9	12,9	16,1
	2	10	32,3	32,3	48,4
	3	9	29,0	29,0	77,4
	4	7	22,6	22,6	100,0
	Gesamt	31	100,0	100,0	

Diese Fehlerquote kann als erfreulich niedrig eingestuft werden.

4.6 Suche nach Daten

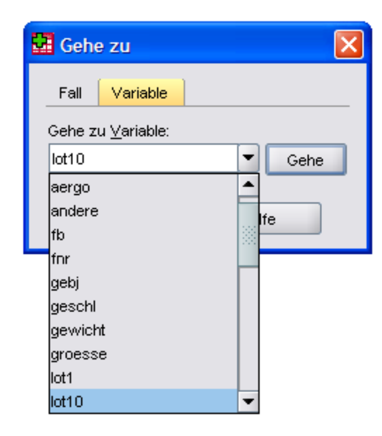
In der Häufigkeitstabelle zu LOT10 haben wir den unzulässigen Wert Null (mit Häufigkeit Eins) entdeckt. Nun möchten wir natürlich wissen, bei welchem Fall dieser Wert auftritt, um eine Korrektur vornehmen zu können. Der betroffene Fall ist leicht zu ermitteln:

- Holen Sie nötigenfalls das Datenfenster in den Vordergrund.
- Markieren Sie in der **Datenansicht** die Variable LOT10 durch einen Klick auf ihren Namen in der Spaltenbeschriftungszone.

In unserem kleinen Datensatz ist eine einzelne Variable leicht zu lokalisieren. SPSS eignet sich aber auch für Projekte mit Tausenden von Variablen und stellt über den Menübefehl

Bearbeiten > Gehe zu Variable...

mit der Dialogbox **Gehe zu** eine sehr nützliche Navigationshilfe zur Verfügung, z.B.:



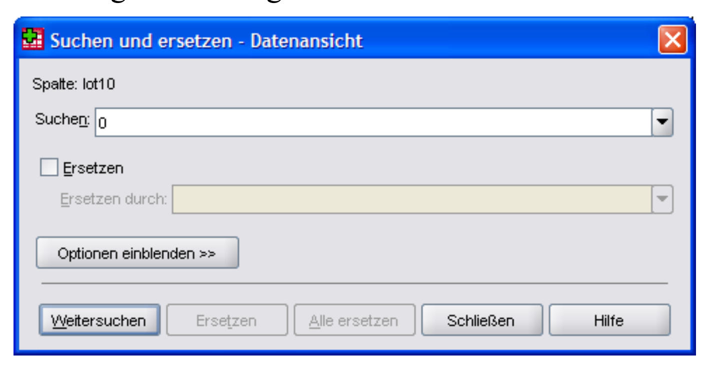
Passend zum bereits eingetippten Namensanfang wird eine Liste mit allen Variablennamen geöffnet und die erste kompatible Zeile markiert. Nach dem Quittieren einer Auswahl ist im Datenfenster die zugehörige Variable markiert.

Im Datenfenster mit der markierten Variablen LOT10 findet man leicht zu den Fällen mit einem interessierenden Wert:

- Klicken Sie auf das Symbol, oder wählen Sie den Menübefehl:

  Bearbeiten > Suchen...

  Dann erscheint die folgende Dialogbox:



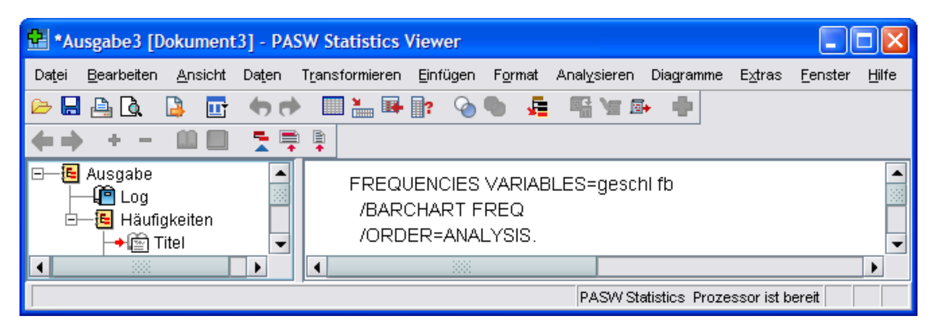
- Tragen Sie den zu suchenden Wert ein, und klicken Sie auf den Schalter **Weitersuchen**. Für die Suche nach SYSMIS ist ein Punkt einzutragen.
- Daraufhin markiert SPSS die erste Trefferzelle, und Sie kennen den Fall mit fehlerhaftem LOT10-Wert: Es ist zufällig der erste Fall (FNR = 1), dessen ausgefüllter Fragebogen im Manuskript wiedergegeben ist (siehe Seite 27), so dass Sie den korrekten Wert ablesen und im Datenfenster eintragen können. Nach dieser Datenkorrektur sollten Sie die Arbeitsdatei sichern und damit die SPSS-Datendatei **kfar.sav** auf den neuen Stand bringen.

5 Speichern der SPSS-Kommandos zu wichtigen Anweisungsfolgen

5.1 Zur Motivation

Eventuell erhalten Sie nach Abschluss der Fehlerkontrolle noch weitere bearbeitete Fragebögen. Sie freuen sich natürlich über die Stichprobenerweiterung und erfassen sofort die neuen Fälle. Dann allerdings fällt Ihnen ein, dass nun alle Kontrollanalysen wiederholt werden müssen.

Um solchen Frust zu vermeiden, brauchen wir eine Möglichkeit, aufwändige und potentiell mehrfach benötigte Anweisungssequenzen zur späteren Wiederverwendung abzuspeichern. In SPSS eignen sich dazu die **Kommandos**, die den einzelnen Dialogboxen zugrunde liegen, und die von SPSS stets im Hintergrund erzeugt und ausgeführt werden, wenn wir eine ausgefüllte Dialogbox mit **OK** abschicken. Wie Sie inzwischen wissen, werden diese Kommandos per Voreinstellung im Ausgabefenster protokolliert, z.B. bei der Häufigkeitsanalyse für die Variablen GESCHL und FB:



In diesem Zusammenhang lohnt ein kurzer Blick auf die Architektur des SPSS-Systems, das aus den beiden folgenden Komponenten besteht:

- **Bedienoberfläche**
  Wir interagieren mit der Bedienoberfläche, die unsere Anweisungen entgegennimmt und die Ergebnisse präsentiert. Wir können der Bedienoberfläche unsere Anweisungen in Form von ausgefüllten Dialogboxen oder als Folge von SPSS-Kommandos übergeben.
- **SPSS-Prozessor**
  Die Bedienoberfläche gibt unsere Anweisungen in jedem Fall in Form von SPSS-Kommandos an den Prozessor weiter, der im Hintergrund arbeitet. Wir erfahren übrigens in der Statuszeile der SPSS-Fenster, was der Prozessor gerade treibt. Da wir den Prozessor bislang nur minimal belastet haben, war in der Statuszeile meist zu lesen:

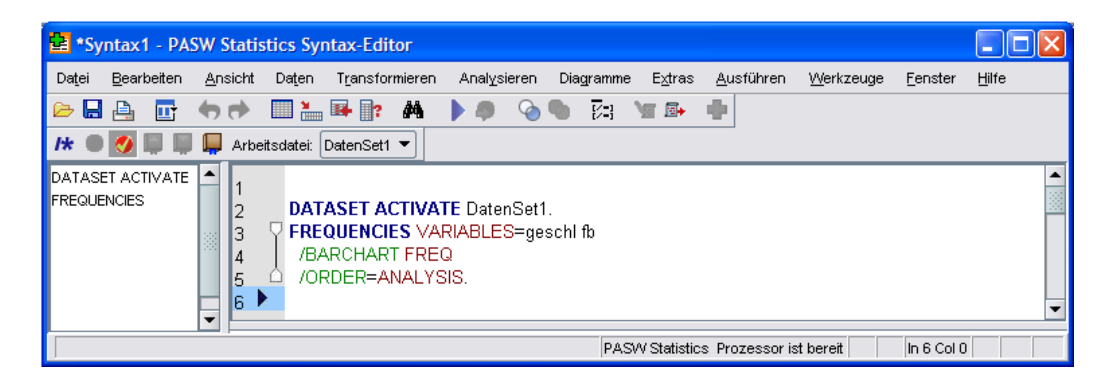
  PASW Statistics Prozessor ist bereit

  Während der Prozessor arbeitet, wird in der Statuszeile angezeigt, mit welchem SPSS-Kommando er gerade beschäftigt ist. Nach dem Abschicken einer Häufigkeitsdialogbox erscheint z.B. (bei unserem kleinen Datensatz allerdings nur sehr kurz):

  Ausführen: FREQUENCIES

  Wenn wir eine ausgefüllte Häufigkeitsdialogbox mit **OK** quittieren, führt der SPSS-Prozessor also im Hintergrund das korrespondierende FREQUENCIES-Kommando aus.

In fast allen SPSS-Dialogboxen kann man über die Standardschaltfläche **Einfügen** die zugrunde liegenden SPSS-Kommandos produzieren lassen. Diese werden dann *nicht* ausgeführt, sondern in ein so genanntes **Syntaxfenster** übertragen, das die Bearbeitung von Kommandos über Zeilennummern, farbliche Unterscheidung verschiedener Syntaxbestandteile und eine intelligente Syntaxvervollständigung unterstützt:



Hier kann man zusammen gehörige Kommandos zu einer Sequenz ansammeln, nach Bedarf einzeln oder geschlossen ausführen lassen und schließlich in einer Datei abspeichern. Später kann man die Kommandos aus dieser Datei wieder laden und, eventuell nach manueller Überarbeitung, erneut ausführen lassen. Das genaue Vorgehen wird in Abschnitt 5.2 an einem konkreten Beispiel geübt.

Eine Folge von SPSS-Kommandos kann man (leicht hochstaplerisch) als **SPSS-Programm** bezeichnen. In fast jedem Projekt sollte es mindestens *ein* SPSS-Programm geben, nämlich das bereits in Abschnitt 3.2.6 vorgeschlagene **Transformationsprogramm**, das aus der Rohdatendatei durch diverse Transformationen die Fertigdatendatei des Projekts erstellt. Wir werden für unser KFA-Projekt ein solches Programm in Abschnitt 0 erstellen.

Ob sich bei einer konkreten Anweisungssequenz das Abspeichern als SPSS-Programm lohnt, muss von Fall zu Fall entschieden werden. Bei kurzen, simplen Sequenzen mit geringer Wiederholungswahrscheinlichkeit ist ein Konservieren unrentabel.

Es soll nicht verschwiegen werden, dass die Ausführung einer Anweisungssequenz mit dem Umweg über ein Syntaxfenster geringfügig mehr SPSS-Kenntnisse erfordert als die direkte Ausführung durch Quittieren von Dialogboxen mit **OK**. Wer sich beim Umgang mit SPSS-Kommandos unsicher fühlt, bei seinem relativ kleinen Projekt eventuell erforderliche Wiederholungen von Dialogbox-Sequenzen nicht scheut und das Risiko inkonsistenter Datenzustände durch große Sorgfalt kontrolliert, der kann auf das Erzeugen und Abspeichern von SPSS-Kommandos verzichten.

Für ambitionierte SPSS-Anwender(innen) muss noch klargestellt werden, dass die Erstellung, Überarbeitung und Ausführung von Programmen in einem Syntaxfenster eine eigenständige Methode der SPSS-Benutzung darstellt, über die fast alle Leistungen des Programms erreichbar sind. Viele SPSS-Optionen stehen sogar *ausschließlich* über die Syntax zur Verfügung, z.B.:

- Die Conjoint-Analyse
- Kontrollstrukturen, mit denen man komplexe Datentransformationen auf effiziente Weise durchführen kann (wie z.B. die DO REPEAT - Schleife)
- Die MATRIX-Programmiersprache, mit der man eigene Statistikprozeduren erstellen kann

Im aktuellen Abschnitt 5 werden nur sehr elementare Hinweise zur Kommandosprache gegeben. Diese sollten genügen für Anwender, die nicht frei programmieren, sondern nur gelegentlich ein von SPSS automatisch erzeugtes Kommando modifizieren wollen. Der Anhang enthält eine ausführlichere Beschreibung der Kommandosprache. Eine vollständige Dokumentation auf ca. 2400 Seiten finden Sie als PDF-Dokument im Hilfesystem von SPSS über

Hilfe > Befehlssyntax-Referenz

Wie schon erwähnt, sind die Dialogboxen beim Erstellen eines SPSS-Programms sehr nützlich. Mit Hilfe der bislang ignorierten Standardschaltfläche **Einfügen** kann die zu einer Dialogbox-Bearbeitung äquivalente Kommandofolge in ein Syntaxfenster übertragen werden. Sie müssen sich also nicht zwischen zwei unabhängigen SPSS-Bediensystemen entscheiden, sondern sollten eine rationelle Kombination der beiden Techniken verwenden.

5.2 Dialogunterstützte Erstellung von SPSS-Programmen

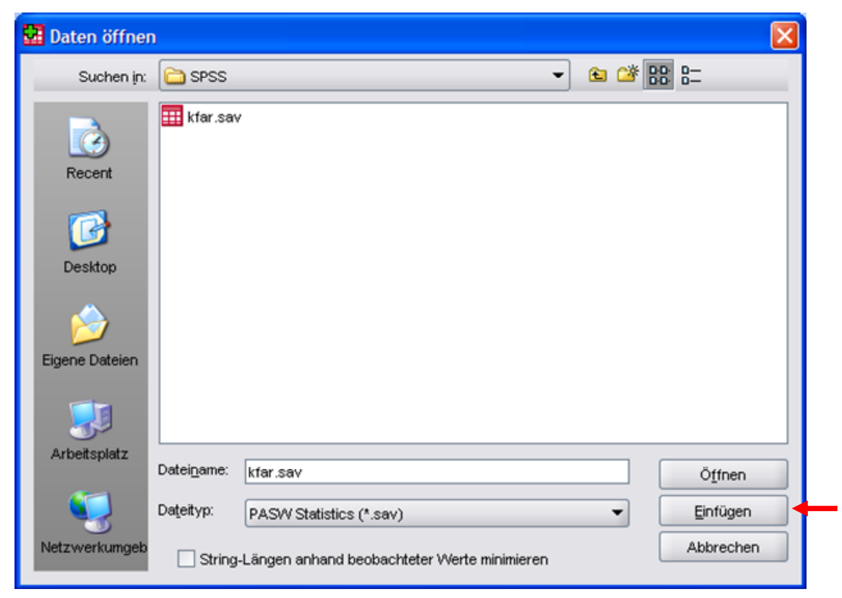
Das folgende SPSS-Programm führt für unser KFA-Projekt die Verteilungs- und Fehleranalysen bei den Variablen GESCHL, FB, GEBJ, GROESSE, GEWICHT, AERGO und AERGM durch (vgl. Abschnitt 4):

```
GET
  FILE='U:\Eigene Dateien\SPSS\kfar.sav'.
DATASET NAME DatenSet1 WINDOW=FRONT.
DATASET ACTIVATE DatenSet1.
FREQUENCIES VARIABLES=geschl fb
  /BARCHART FREQ
  /ORDER=ANALYSIS.
FREQUENCIES VARIABLES=gebj groesse gewicht aergo aergm
  /STATISTICS=STDDEV VARIANCE MINIMUM MAXIMUM
   MEAN MEDIAN MODE
   SKEWNESS SESKEW KURTOSIS SEKURT
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.
```

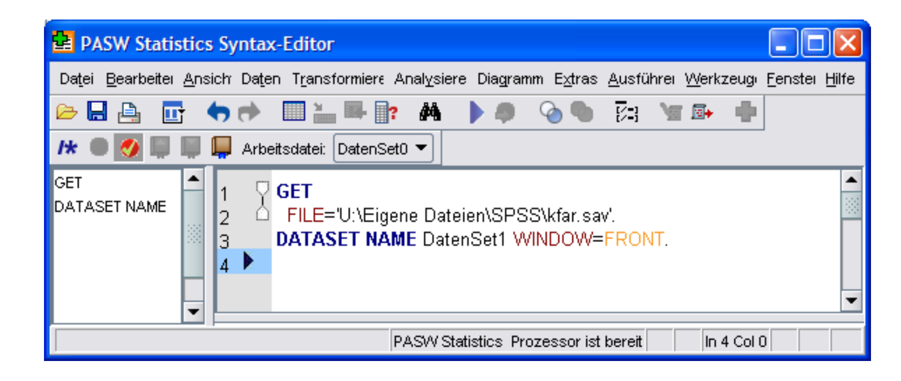
Wir werden dieses Programm gleich „vollautomatisch“ mit drei Mausklicks auf **Einfügen**-Schalter produzieren und dabei auch seine Bestandteile kurz beschreiben. Als Ausgangssituation für die anschließenden Erläuterungen wird eine neue SPSS-Sitzung mit einem leeren Datenfenster angenommen. Verzichten Sie also beim SPSS-Start auf das Öffnen einer Datendatei per Startassistent (z.B. durch einen Mausklick auf den Schalter **Abbrechen**). Dabei erhalten Sie ein leeres Datenfenster mit dem Namen **DatenSet0**. Rufen Sie die Dialogbox zum Öffnen einer Datendatei mit dem folgenden Menübefehl auf:

Datei > Öffnen > Daten

Navigieren Sie zum Ordner mit Ihrer Rohdatendatei, schreiben oder klicken Sie deren Namen in das Feld **Dateiname**, und betätigen Sie dann den Schalter **Einfügen**.



Daraufhin beginnt SPSS *nicht* damit, aus der angegebenen Datendatei ein neues Datenset zu erstellen und zur Arbeitsdatei zu machen, sondern schreibt das für diese Aktionen zuständige GET-Kommando in ein Syntaxfenster mit dem Titel **Syntax1**:



Das GET-Kommandos ist sehr einfach aufgebaut:

- Es beginnt mit dem Kommandonamen GET.
- Im FILE-Subkommando wird die zu öffnende Datendatei angegeben.
- Am Ende muss wie bei jedem SPSS-Kommando ein **Punkt** stehen.

Das zusätzlich erzeugte Kommando DATASET NAME hat bei der Ausführung folgende Effekte:

- Das aktive Datenblatt (die Arbeitsdatei) erhält einen neuen Namen.
- Das beteiligte Dateneditorfenster wird in den Vordergrund geholt.

Noch sind die beiden Kommandos nicht ausgeführt worden, so dass die aktuelle Arbeitsdatei (das **DatenSet0**) noch leer ist. Der Versuch, zur Produktion des ersten FREQUENCIES-Kommandos mit dem Menübefehl

Analysieren > Deskriptive Statistik > Häufigkeiten...

eine Häufigkeits-Dialogbox zu öffnen, führt zur Meldung:



Daher wollen wir jetzt die Kommandos GET und DATASET NAME ausführen lassen, um die Daten einzulesen. Wählen Sie dazu im Syntaxfenster den Menübefehl

Ausführen > Alles

Daraufhin erstellt SPSS ein neues Datenblatt mit den Rohdaten, das

- zur Arbeitsdatei (zum aktiven Datenblatt) wird,
- mit der Rohdatendatei **kfar.sav** verbunden ist,
- den Namen **DatenSet1** erhält
- und in den Vordergrund gelangt.

Das beim Programmstart angebotene und nun überflüssig gewordene leere **DatenSet0** wird automatisch geschlossen.

Spezifizieren Sie jetzt mit Hilfe der zuständigen Dialogbox dieselbe Häufigkeitsanalyse zu den Variablen GESCHL und FB wie in Abschnitt 4.3. Verlassen Sie die Dialogbox jedoch nicht mit **OK**, sondern mit **Einfügen**. Daraufhin erscheint am Ende des Syntaxfensters ein FREQUENCIES-Kommando (siehe oben):

- Es beginnt mit dem Kommandonamen FREQUENCIES.
- Im VARIABLES-Subkommando ist angegeben, welche Variablen analysiert werden sollen.
- Das BARCHART-Subkommando sorgt dafür, dass Balkendiagramme mit einer Häufigkeitsbeschriftung erscheinen.
- Das ORDER-Subkommando entscheidet bei der Analyse mehrerer Variablen darüber, ob die Statistiken für jede Variable in einer eigenen Tabelle oder für alle Variablen in einer gemeinsamen Tabelle erscheinen sollen. Um diese Entscheidung in der **Häufigkeiten**-Dialogbox zu treffen, müssen Sie übrigens die **Format**-Subdialogbox öffnen und im Rahmen **Mehrere Variablen** die passende Option wählen.
- Das FREQUENCIES-Kommando wird wie jedes SPSS-Kommando durch einen **Punkt** abgeschlossen.

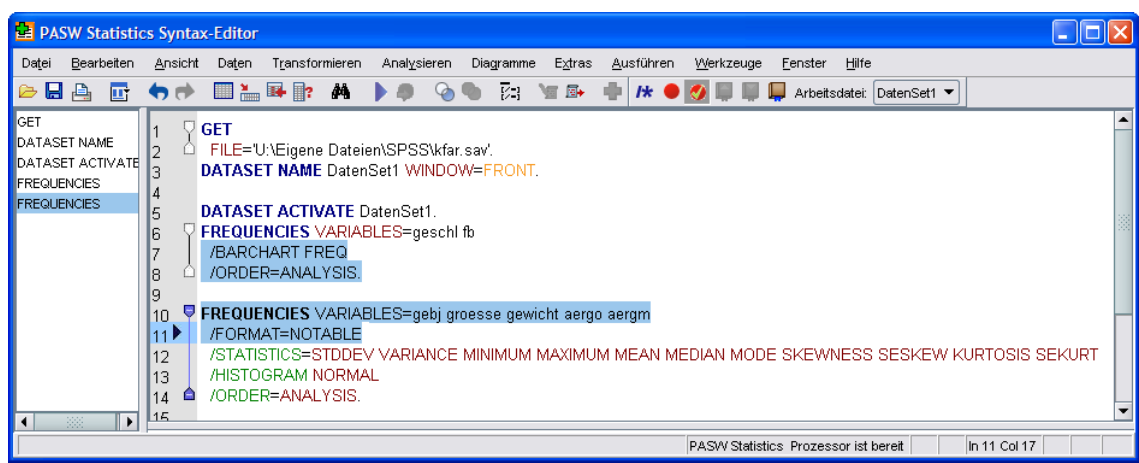
Produzieren Sie als Nächstes die Syntax zu der in Abschnitt 4.5 durchgeführten Häufigkeitsanalyse für die Variablen GEBJ, GROESSE, GEWICHT, AERGO und AERGM.

Nun sollte Ihr Syntaxfenster den zu Beginn des Abschnitts wiedergegebenen Inhalt haben. Die Kommandos GET und DATASET NAME sind schon gelaufen. Das zusammen mit der ersten Häufigkeitsanalyse automatisch erstellte Kommando

```
DATASET ACTIVATE DatenSet1.
```

aktiviert vorsichtshalber das mit der Rohdatendatei **kfar.sav** verbundene **DatenSet1**. Es ist momentan überflüssig, weil gar kein anderes Datenset existiert, kann aber bedenkenlos ausgeführt werden. Um die Häufigkeitsanalysen zu erhalten, müssen Sie noch die beiden FREQUENCIES-Kommandos ausführen lassen. Weil es sich um eine Teilmenge der im Syntaxfenster vorhandenen Kommandos handelt, müssen Sie folgendermaßen vorgehen:

- Markieren Sie zunächst per Maus *die beiden* auszuführenden Kommandos, wobei von jedem Kommando wenigstens ein Zeichen in die Markierung einbezogen werden muss, z.B.:



- Klicken Sie dann auf den Symbolleistenschalter ▶, oder drücken Sie die Tastenkombination **Strg+R**. Daraufhin werden alle Kommandos im Syntaxfenster ausgeführt, die (zumindest teilweise) markiert sind.

Im Ausgabefenster protokolliert SPSS per Voreinstellung die verarbeiteten Kommandos in **Log**-Teilausgaben, falls Sie dieses Verhalten nicht per **Optionen**-Dialog auf der Registerkarte **Viewer** abschalten (siehe Abschnitt 3.2.5). Außerdem protokolliert SPSS zu jeder Analyseanforderung in der zunächst zugeklappten Teilausgabe **Anmerkungen** u.a. die zugrunde liegende Syntax, z.B.:

Anmerkungen

Ausgabe erstellt		03-Jun-2009 03:22:31
Kommentare		
Eingabe	Daten	U:\Eigene Dateien\SPSS\kfar.sav
	Aktiver Datensatz	DatenSet1
	Filter	<keine>
	Gewichtung	<keine>
	Aufgeteilte Datei	<keine>
	Anzahl der Zeilen in der Arbeitsdatei	31
Behandlung fehlender Werte	Definition von fehlenden Werten	Benutzerdefinierte fehlende Werte werden als fehlend behandelt.
	Verwendete Fälle	Statistik basiert auf allen Fällen mit gültigen Daten.
Syntax		FREQUENCIES VARIABLES=geschl fb /BARCHART FREQ /ORDER=ANALYSIS.
Ressourcen	Prozessorzeit	0:00:01.641
	Verstrichene Zeit	0:00:01.625

Damit sich durch spätere Wiederverwendung der SPSS-Kommandos der gewünschte Rationalisierungseffekt einstellen kann, müssen Sie Ihr SPSS-Programm sichern. Wechseln Sie dazu nötigenfalls zum Syntaxfenster, und wählen Sie den Menübefehl:

Datei > Speichen unter...

Verwenden Sie im Dateinamen die vorgeschlagene Erweiterung **sps**, indem Sie *keine* Erweiterung angeben.

Wenn Sie später dieselbe Auswertung nochmals benötigen, müssen Sie lediglich das vorhandene Programm mit dem Menübefehl:

Datei > Öffnen > Syntax

laden und ausführen lassen.

Um die Ausführung *sämtlicher* Kommandos in einem Syntaxfenster anzuordnen, haben Sie folgende Möglichkeiten:

- Menübefehl **Ausführen > Alles**
- Alle Kommandos markieren (z.B. mit **Strg+A**) und die Ausführung anfordern (z.B. per Mausklick auf das Symbol oder mit der Tastenkombination **Strg+R**)

Lässt man obiges Programm innerhalb einer SPSS-Sitzung erneut komplett ausführen, erscheint die folgende Warnung im Ausgabefenster:

Warnungen

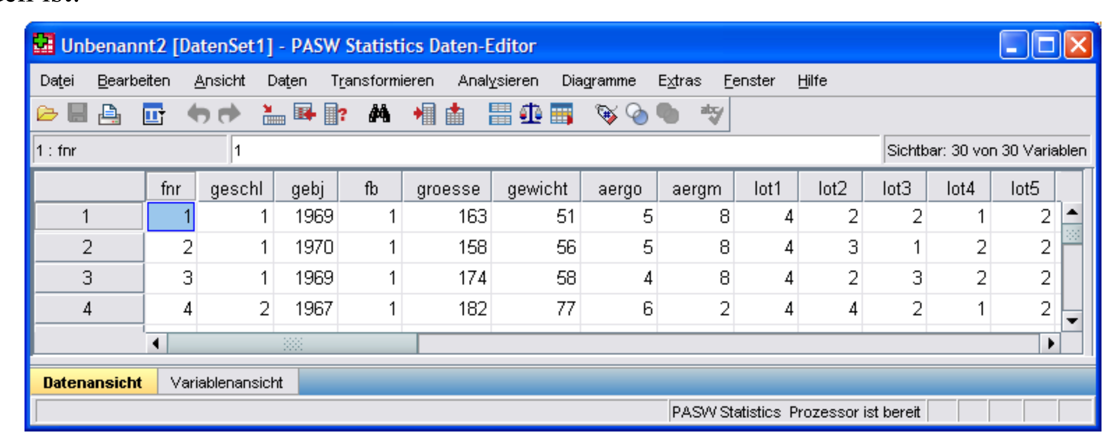
Das aktive Daten-Set ersetzt das vorhandene Daten-Set mit dem Namen DatenSet1.

Leider sorgt die generell begrüßenswerte Möglichkeit, in einer SPSS-Sitzung *mehrere* Datenblätter zu verwenden, aktuell für eine Komplikation. Damit daraus keine Konfusion wird, müssen wir das Geschehen im Detail verfolgen:

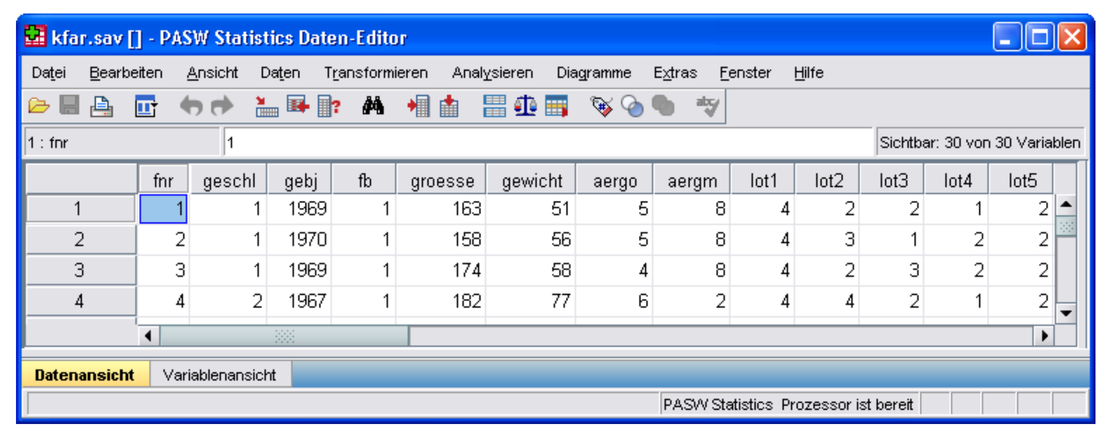
- Vor dem erneuten Ausführen des Programms hatte die Arbeitsdatei (das aktive Datenblatt) den Namen **DatenSet1** und war mit der Rohdatendatei **kfar.sav** verbunden.
- Das erneut ausgeführte GET-Kommando erzeugt ein neues Datenblatt, kopiert den Inhalt der Rohdatendatei dorthin und aktiviert das neue Datenblatt (macht es zur Arbeitsdatei).
- Die Rohdatendatei bleibt aber mit dem älteren Datenblatt verbunden, das noch den Namen **DatenSet1** trägt.

- Das erneut ausgeführte Kommando DATASET NAME gibt dem aktiven Datenblatt (der aktuellen Arbeitsdatei) den bereits in Verwendung befindlichen Namen **DatenSet1**, woraufhin das alte Datenset mit diesem Namen geschlossen wird (siehe Warnung).

Insgesamt führt die erneute Ausführung des Programms dazu, dass ein Datenblatt namens **DatenSet1** mit dem Inhalt der Rohdatendatei existiert, das aber *nicht* mit der Rohdatendatei verbunden ist:



Wenn dieses Verhalten stört, kann man z.B. die Kommandos DATASET NAME und DATASET ACTIVATE streichen. Dann bleibt das per GET befüllte und mit der Rohdatendatei verbundene Datenset unbenannt und wird bei jeder Ausführung des Programms überschrieben:



5.3 *Arbeiten mit dem Syntax-Fenster*

Im Syntaxfenster lassen sich automatisch erstellte SPSS-Kommandos leicht modifizieren, um z.B. die in einer Statistikprozedur zu analysierenden Variablen auszutauschen.

Man kann ein neues Syntaxfenster auch unabhängig vom **Einfügen**-Schalter einer Dialogbox direkt anfordern mit:

Datei > Neu > Syntax

Wenn *mehrere* Syntaxfenster vorhanden sind, muss geregelt werden, in welches Fenster SPSS die per **Einfügen**-Schalter automatisch erzeugten Kommandos übertragen soll. Dies geschieht genauso wie bei den Ausgabefenstern: Ein Mausklick auf den aktiven Schalter in seiner Symbolleiste macht ein Syntaxfenster zum **Hauptfenster** in seiner Kategorie. Es ist an einem Pluszeichen im Symbol zum Systemmenü zu erkennen (siehe linken Rand der Titelzeile).

Um ein bestimmtes Syntaxfenster in den Vordergrund zu holen, können Sie es anklicken oder das **Fenster**-Menü eines beliebigen SPSS-Fensters benutzen. Jedes Syntaxfenster kann auf

Windows-übliche Weise geschlossen werden, z.B. indem Sie es in den Vordergrund holen und dann anordnen:

Datei > Schließen

Wenn Sie längere Zeit mit SPSS arbeiten, wird sich vermutlich Ihr Umgang mit SPSS-Syntax in folgenden Stufen weiterentwickeln:

- Kommandos automatisch erzeugen lassen und später unverändert wiederverwenden
  Bei dieser Arbeitsweise müssen Sie nur wissen, wie man SPSS-Kommandos per Dialogbox in ein Syntaxfenster befördert, und wie man überflüssige Kommandos löscht.
- Automatisch erzeugte Kommandos modifizieren
  Es zeigt sich, dass SPSS-Kommandos meist leicht zu durchschauen und zu modifizieren sind.
- Freies Programmieren

5.4 Elementare Regeln zur SPSS-Syntax

Für den im Kurs vorgeschlagenen Einsatz von SPSS-Kommandos sollte die Kenntnis der folgenden Regeln genügen:

- Ein Kommando besteht aus seinem Namen und den Spezifikationen, die sich aus Schlüsselwörtern (z.B. VARIABLES, STATISTICS), Variablennamen usw. zusammensetzen, z.B.:

```
Kommandoname  ⟶ FREQUENCIES
                  VARIABLES=geschl fb
Spezifikationen { /BARCHART FREQ
                  /ORDER=ANALYSIS.
```

- Zwei Elemente der Kommandosprache sind durch mindestens ein Leerzeichen oder durch einen Zeilenwechsel voneinander zu trennen. Manche Zeichen mit festgelegter Bedeutung wie z.B. "=", "/", "(", "+", ">" sind aber selbstbegrenzend, d.h. davor und danach sind keine Leerzeichen nötig (aber erlaubt).
- Ein Kommando kann sich über beliebig viele Fortsetzungszeilen erstrecken, dabei dürfen aber *innerhalb* des Kommandos keine Leerzeilen auftreten. Diese signalisieren nämlich per Voreinstellung (wie der Punkt) das Ende des Kommandos.
- Zwischen zwei Kommandos dürfen beliebig viele Leerzeilen stehen, was eine übersichtliche Gestaltung von SPSS-Programmen erlaubt.
- **Jedes Kommando muss in einer neuen Zeile beginnen und mit einem Punkt enden.**

Gut kommentierte Programme sind später leichter zu verstehen. Die SPSS-Syntax bietet zum Kommentieren das Kommando COMMENT, dessen Name durch ein Sternchen ersetzt werden darf, z.B.:

```
* Mit diesem Programm wird die Rohdatendatei KFAR.SAV
  auf Erfassungsfehler untersucht.
GET
  FILE='U:\Eigene Dateien\SPSS\KFAR.SAV'.
   .   .   .
```

Beachten Sie beim Kommentar-Kommando:

- Es darf sich über beliebig viele Fortsetzungszeilen erstrecken, wobei innerhalb des Kommandos keine Leerzeilen erlaubt sind.
- **Jedes Kommentar-Kommando muss mit einem Punkt abgeschlossen werden.** Wenn Sie den Punkt am Ende vergessen, dann betrachtet SPSS den folgenden Programmtext bis zum nächsten Punkt (oder zur nächsten Leerzeile) als Teil des Kommentars!
- Endet eine Kommentarzeile mit einem Punkt, so betrachtet SPSS das Kommentar-Kommando als abgeschlossen. Wenn Sie einen Punkt als *Satzzeichen* ans Ende einer Kommentarzeile gesetzt haben, dann müssen Sie die nächste Kommentarzeile wieder mit COMMENT oder ***** einleiten.
- Punkte innerhalb einer Kommentarzeile sind kein Problem.

6 Datentransformation

6.1 Vorbemerkungen

Die zur Untersuchung unserer differentialpsychologischen Hypothese benötigte Optimismus-Variable existiert noch nicht, sondern muss erst aus den 12 LOT-Variablen berechnet werden. Vor dieser Berechnung müssen allerdings die aus messtechnischen Gründen umgepolten (negativ formulierten) LOT-Fragen geeignet rekodiert werden (z.B. Frage 3).[1] Es ist typisch für empirische Studien, dass vor dem eigentlichen Start der Datenanalyse aus den Rohvariablen mit zahlreichen Datentransformationen neue oder modifizierte Fertigvariablen erstellt werden müssen. Dabei geht es sowohl um sorgfältig zu absolvierende Pflichtübungen als auch um kreative Begriffsbildungen mit dem Ziel, durch geschickte Kombination vorhandener Informationen begrifflichen Mehrwert zu schaffen. Wir werden z.B. aus den „einfachen" Begriffen *Gewicht* und *Größe* den für unsere ernährungswissenschaftlichen Begleitstudien relevanten *Body Mass Index* nach folgender Formel berechnen:

$$\frac{Gewicht\ (in\ kg)}{Größe^2\ (in\ m)}$$

In diesem Abschnitt werden Sie häufig benötigte SPSS-Befehle zur Datentransformation kennen lernen. Diese wirken sich auf die Arbeitsdatei (das aktive Datenblatt) aus, wo entweder neue Variablen aufgenommen oder vorhandene Variablen verändert werden.

Per Voreinstellung werden dabei *alle Fälle gleichermaßen* behandelt. Man kann die Ausführung einer Datentransformation aber auch von einer Bedingung abhängig machen, so dass nicht mehr alle Fälle davon betroffen sind. Diese Möglichkeit werden wir dazu verwenden, die MD-Behandlung bei den Motiv-Variablen in Ordnung zu bringen, indem wir genau für die Fälle mit

MOTIV1 = MOTIV2 = ... = ANDERE = 0

bei allen genannten Variablen die Null durch SYSMIS ersetzen.

SPSS unterstützt Transformationen für Variablen beliebigen Typs. Wir beschränken uns jedoch auf die besonders wichtigen numerischer Variablen.

6.1.1 Rohdatendatei, Transformationsprogramm und Fertigdatendatei

In Abschnitt 3.2.6 wurde vorgeschlagen, zu jedem Projekt ein SPSS-Transformationsprogramm zu erstellen, dessen Aufgabe darin besteht, ausgehend von der Rohdatendatei alle Fertigvariablen zu entwickeln, die im weiteren Verlauf routinemäßig benötigt werden. *Alle* potentiell relevanten Variablen (roh oder fertig) sollen in eine erweiterte Datendatei gesichert werden, die sich für alle Auswertungsarbeiten eignet.[2] Mit Rücksicht auf diese Idee haben wir die bislang existierende Datendatei mit **kfar.sav** (*r* für *roh*) bezeichnet. Im Namen der Fertigdatendatei werden wir das **r** dann weglassen.

Wir werden im Verlauf des aktuellen Abschnitts 6 das SPSS-Transformationsprogramm zu unserem KFA-Projekt erstellen, indem wir passend konfigurierte Dialogboxen mit dem Schalter **Ein-**

---

[1] Das Rekodieren ist keine zwingende Voraussetzung für die Berechnung des Optimismus-Schätzwerts, hat aber gravierende Vorteile (z.B. einfachere Berechnung, Möglichkeit zur Skalenanalyse).

[2] Unter gewissen, am ehesten in großen Projekten anzutreffenden Umständen kann es sinnvoll bzw. notwendig sein, die auszuwertenden Daten in *mehreren* Dateien bereitzuhalten. Werden die Variablen oder Fälle einer Tabelle auf mehrere Dateien verteilt, kann es leicht zu dem Problem kommen, dass sich die in einer Analyse zu vergleichenden Fälle oder Variablen in verschiedenen Dateien befinden. Treten in einem Projekt mehrere Tabellen auf (z.B. mit Kunden bzw. Mitarbeitern), werden natürlich entsprechend viele Datendateien benötigt.

fügen quittieren, um die äquivalenten SPSS-Kommandos in einem Syntaxfenster zu sammeln (siehe Abschnitt 5). Dabei ist eine hohe Sorgfalt erforderlich, weil fehlerhafte Anweisungen im Transformationsprogramm schwerwiegende Konsequenzen für die weitere Arbeit haben können.

Das fertige Transformationsprogramm wird anschließend ausgeführt, wobei die Fertigdatendatei entsteht. Außerdem wird das Transformationsprogramm in einer Datei gespeichert, damit es z.B. nach einer Stichprobenerweiterung oder nach einer Fehlerkorrektur in den Rohdaten ohne großen Aufwand erneut ausgeführt werden kann. Als Dateinamen werden wir **kfat.sps** verwenden.

Man kann alle erforderlichen Transformationen auch durch direkte Ausführung der zuständigen Dialogboxen erledigen (Schalter **OK**). Diese Arbeitsweise ist zweifellos für Anfänger leichter zu handhaben als die programmorientierte Methode, hat aber folgende Nachteile:

- Beim sukzessiven manuellen Modifizieren der Datendatei geht bei größeren Projekten leicht der Überblick verloren. Z.B. weiß irgendwann von einer bestimmten Variablen niemand mehr, in welchen Zwischenschritten sie aus welchen anderen Variablen berechnet worden ist. Spätestens nach dem Auftreten unplausibler Ergebnisse muss die *tatsächlich* angewendete Berechnungsvorschrift als mögliche Fehlerquelle überprüft werden. Bei der Verwendung eines Transformationsprogramms ist die Herkunft der abgeleiteten Variablen stets dokumentiert.
- Sind Wiederholungen von Datentransformationen erforderlich, müssen diese komplett neu spezifiziert werden. Solche Wiederholungen sind z.B. nach einer Datenkorrektur fällig, weil SPSS abgeleitete Variablen **nicht** automatisch anpasst, wenn sich Werte der Ursprungsvariablen ändern. Nach Korrekturen bei den Rohvariablen müssen Sie also alle Datentransformationen wiederholen, in die diese Rohvariablen eingehen. Ein weiterer potentieller Anlass für die Wiederholungen von Datentransformationen ist die Erweiterung der Stichprobe.

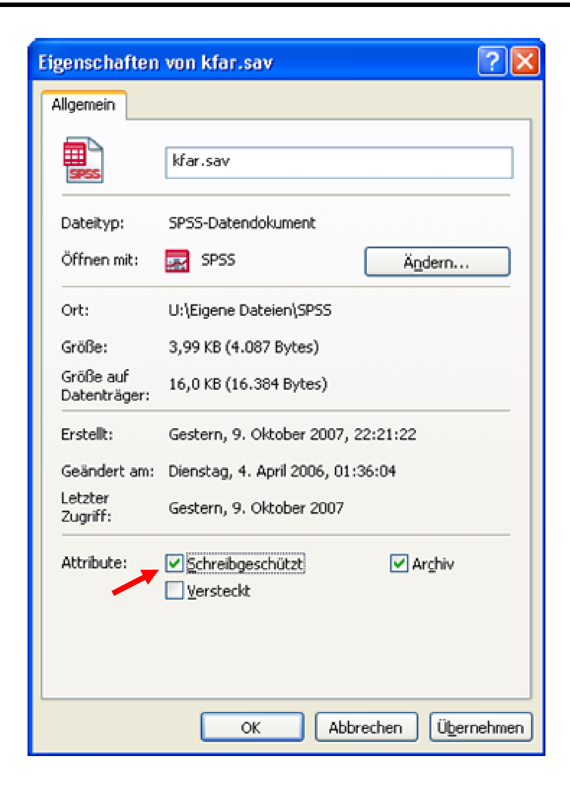
Die für ein Projekt erforderlichen Datentransformationen in Form von SPSS-Kommandos zu konservieren, lohnt sich meistens, denn:

- Die einzelnen Anweisungen sind relativ komplex und damit ebenso fehleranfällig wie zeitaufwändig.
- Es ist relativ wahrscheinlich, dass die gesamte Anweisungsfolge wiederholt durchgeführt werden muss (z.B. bei entdeckten Fehlern in den Rohvariablen oder bei einer Stichprobenerweiterung).
- Die Anweisungen zur Datentransformation sind dokumentationspflichtig.

6.1.2 Hinweise zum Thema Datensicherheit

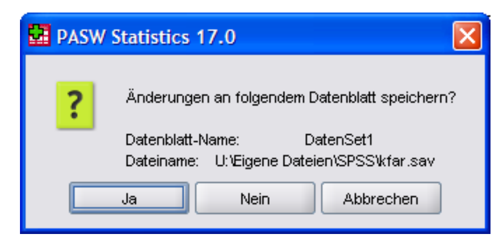
Ihre Rohdaten können nach der sorgfältigen Datenerfassung und -prüfung vorerst als korrekt gelten. Sichern Sie den erreichten Stand, indem Sie die Rohdaten in mindestens **zwei** Dateien speichern (möglichst auf verschiedenen Datenträgern).

Es ist sinnvoll, für beide Dateien das Schreibschutzattribut mit dem Windows-Explorer zu setzen, z.B.:



Vor der geplanten Änderung einer Datei muss das Schreibschutzattribut wieder aufgehoben werden. Ähnlich sorgfältig sollten Sie nach seiner Fertigstellung das Transformationsprogramm aufbewahren.

Wenn Sie beim Verlassen von SPSS gefragt werden, ob Sie das Daten- oder ein Syntaxfenster sichern wollen, sollten Sie sehr sorgfältig prüfen, ob bei dem entsprechenden Objekt während der Sitzung tatsächlich nur geplante Veränderungen stattgefunden haben.



Antworten Sie im Zweifelsfall mit **Nein**. Möglicherweise haben Sie durch unbeabsichtigte Tastendrücke Daten gelöscht oder verändert. Diese Fehler sollten dann auf keinen Fall auf die Festplatte geschrieben werden.

6.1.3 Initialisierung neuer numerischer Variablen

Wenn Sie in einer Datentransformationsanweisung die Erstellung einer *neuen* numerischen Variablen anfordern, dann wird die (Fälle × Variablen) - Datenmatrix in der Arbeitsdatei (im aktiven Datenblatt) um eine Spalte erweitert (am rechten Rand). SPSS **initialisiert** dabei zunächst die neue Variable, indem es für alle Fälle den MD-Indikator SYSMIS als Wert einträgt. Gelingt anschließend die Ermittlung der neuen Variablenausprägung für einen Fall, so wird der Initialwert entsprechend ersetzt. Anderenfalls bleibt SYSMIS stehen, so dass der betroffene Fall bei allen Berechnungen mit der neuen Variablen ausgeschlossen wird.

6.2 Alte Werte einer Variablen auf neue abbilden (Umkodieren)

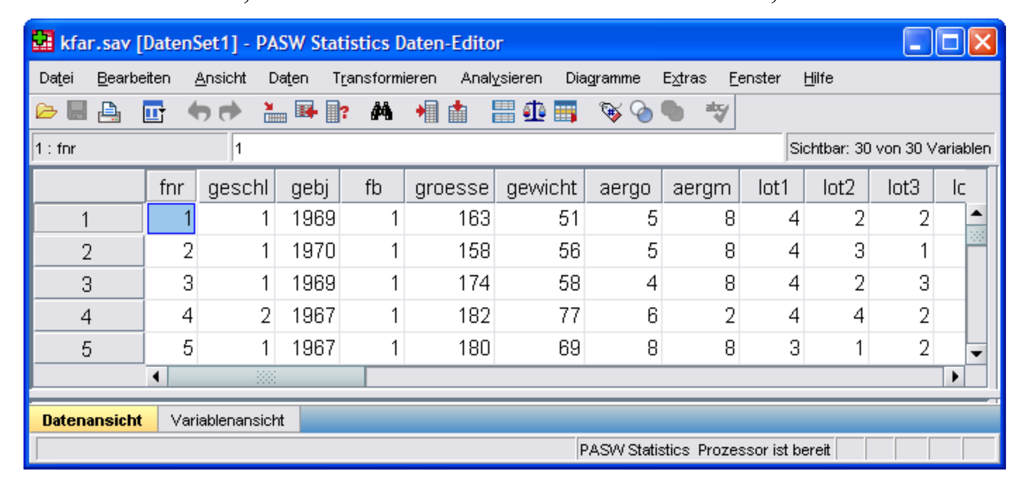
Mit dem Befehl **Umkodieren** aus dem Menü **Transformieren** bzw. mit dem äquivalenten RECODE-Kommando können die Werte einer bestehenden Variablen in neue Werte überführt werden. Man kann die Ausgangsvariable verändern oder eine neue Variable mit dem rekodierten Wertevektor erstellen.

6.2.1 Das praktische Vorgehen am Beispiel einer Gruppenbildung

Da wir im Abschnitt 6 das KFA-Transformationsprogramm sukzessive aufbauen wollen, wird eine Arbeitsdatei mit unseren Rohdaten benötigt. Öffnen Sie daher nötigenfalls über den Menübefehl

Datei > Öffnen > Daten

die Rohdatendatei **kfar.sav**, wobei ein benanntes Datenblatt entsteht, z.B.:



	fnr	geschl	gebj	fb	groesse	gewicht	aergo	aergm	lot1	lot2	lot3
1	1	1	1969	1	163	51	5	8	4	2	2
2	2	1	1970	1	158	56	5	8	4	3	1
3	3	1	1969	1	174	58	4	8	4	2	3
4	4	2	1967	1	182	77	6	2	4	4	2
5	5	1	1967	1	180	69	8	8	3	1	2

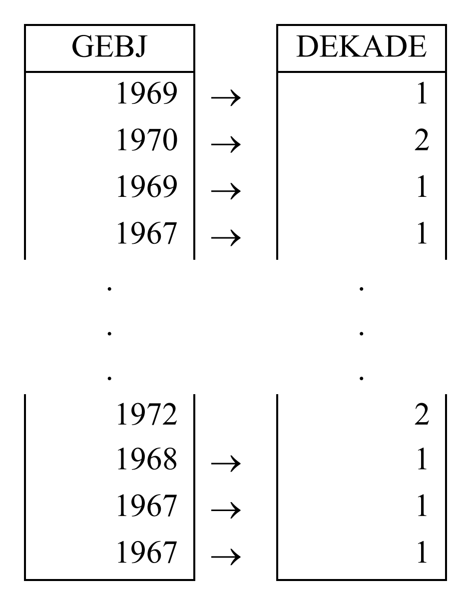
Um das Umkodieren zu üben, wählen wir ein mäßig sinnvolles Beispiel aus unserer Studie: Wir konstruieren unter dem Namen DEKADE eine vergröberte Variante der Jahrgangsvariablen, bei der alle in den 60'er Jahren geborenen Personen den Wert 1 und alle in den 70'er Jahren geborenen Personen den Wert 2 erhalten sollen. Wie man sich anhand der Häufigkeitstabelle zur Variablen GEBJ

Geburtsjahr

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1961	1	3,2	3,2	3,2
	1964	1	3,2	3,2	6,5
	1965	1	3,2	3,2	9,7
	1966	2	6,5	6,5	16,1
	1967	7	22,6	22,6	38,7
	1968	3	9,7	9,7	48,4
	1969	2	6,5	6,5	54,8
	1970	7	22,6	22,6	77,4
	1972	3	9,7	9,7	87,1
	1974	2	6,5	6,5	93,5
	1975	2	6,5	6,5	100,0
	Gesamt	31	100,0	100,0	

überzeugen kann, ist damit für alle Fälle in unserer Stichprobe ein DEKADE-Wert definiert. Mit Hilfe der neuen Variablen kann man z.B. den Einfluss des Geburtsjahrzehnts auf diverse abhängige Variablen untersuchen, wobei man sich von der Informationsreduktion (im Vergleich zu GEBJ) keinen allzu großen Nutzen versprechen sollte.

Bei der geplanten Rekodierung wird die (Fälle × Variablen)-Datenmatrix der Arbeitsdatei um eine neue Variable erweitert, die folgendermaßen aus der vorhandenen Variablen GEBJ entsteht:



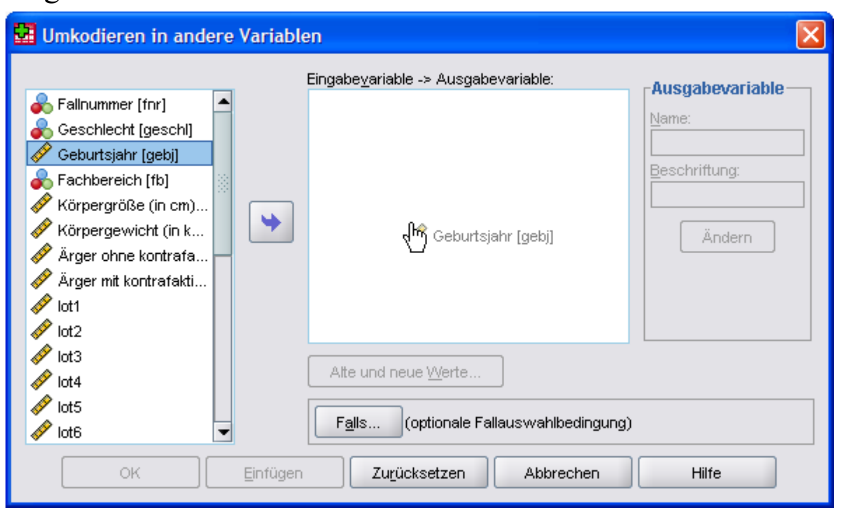
GEBJ		DEKADE
1969	→	1
1970	→	2
1969	→	1
1967	→	1
.		.
.		.
.		.
1972		2
1968	→	1
1967	→	1
1967	→	1

Wählen Sie den Menübefehl:

Transformieren > Umkodieren in andere Variablen

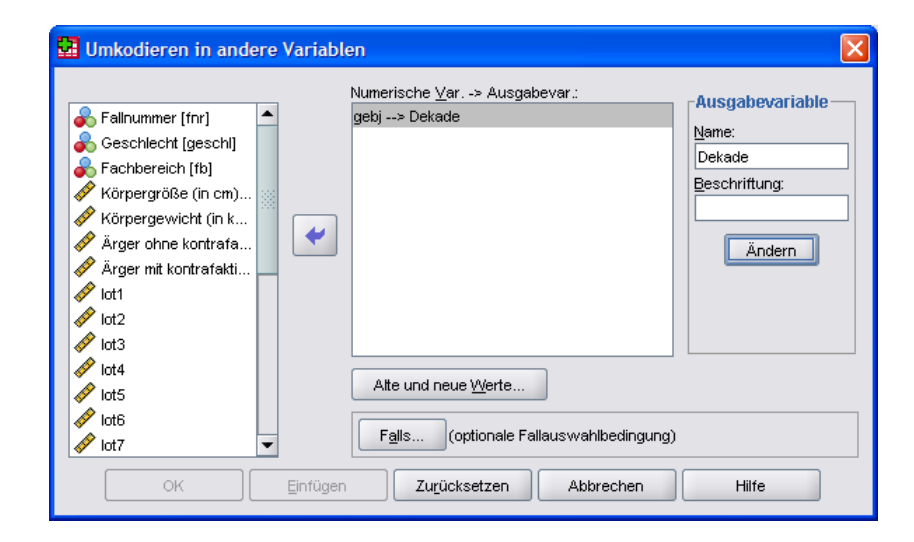
und machen Sie folgendermaßen weiter:

- Befördern Sie in der Dialogbox **Umkodieren in andere Variablen** die Variable GEBJ in das Feld **Eingabevariable -> Ausgabevariable**. Statt den Schalter zu benutzen, können Sie in SPSS solche Transportaufgaben auch per Drag & Drop (Ziehen und Ablegen) erledigen:



- Tragen Sie im Bereich **Ausgabevariable** den gewünschten **Namen** der neu zu erzeugenden Variablen ein.
- Optional kann eine **Beschriftung** (also ein Variablenlabel) ergänzt werden. Wir verzichten darauf, so dass der Variablenname *Dekade* auch zur Beschriftung der Ausgabe verwendet werden wird. In dieser Situation sollte man im Variablennamen auf die korrekte Schreibweise achten.
- Klicken Sie auf **Ändern**.

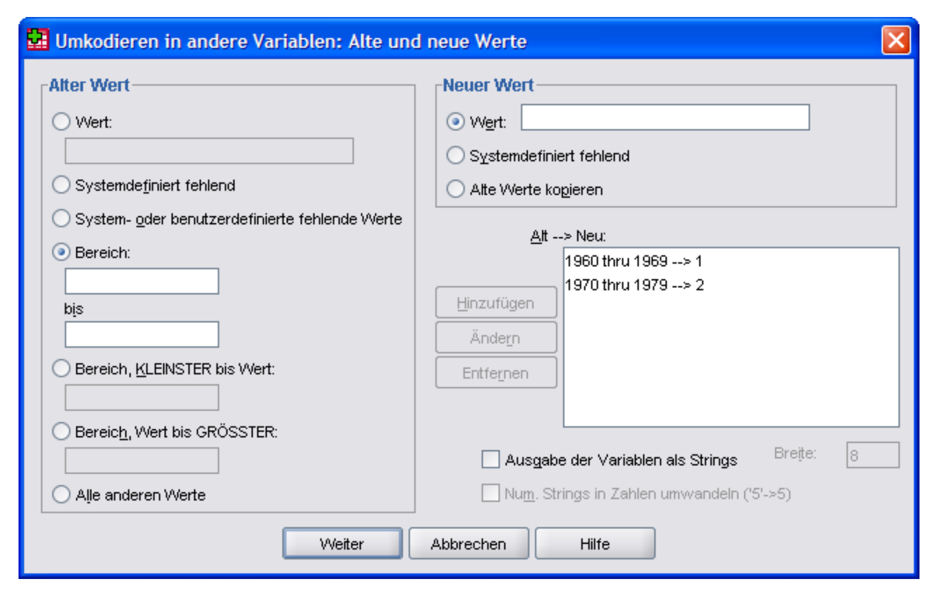
Danach müsste Ihre Dialogbox ungefähr so aussehen:



Legen Sie nun die Abbildungsregeln fest:

- Aktivieren Sie mit dem Schalter **Alte und neue Werte** die Subdialogbox **Umkodieren in andere Variablen: Alte und neue Werte**.
- Geben Sie im Rahmen **Alter Wert** den **Bereich** von 1960 bis 1969 an, und wählen Sie als zugehörigen **Neuen Wert** die Eins.
- Beenden Sie die Definition der ersten Abbildungsvorschrift mit **Hinzufügen**.
- Vereinbaren Sie analog die Zuordnungsvorschrift: „1970 bis 1979 → 2".

Jetzt müssten Sie dieses Bild sehen:

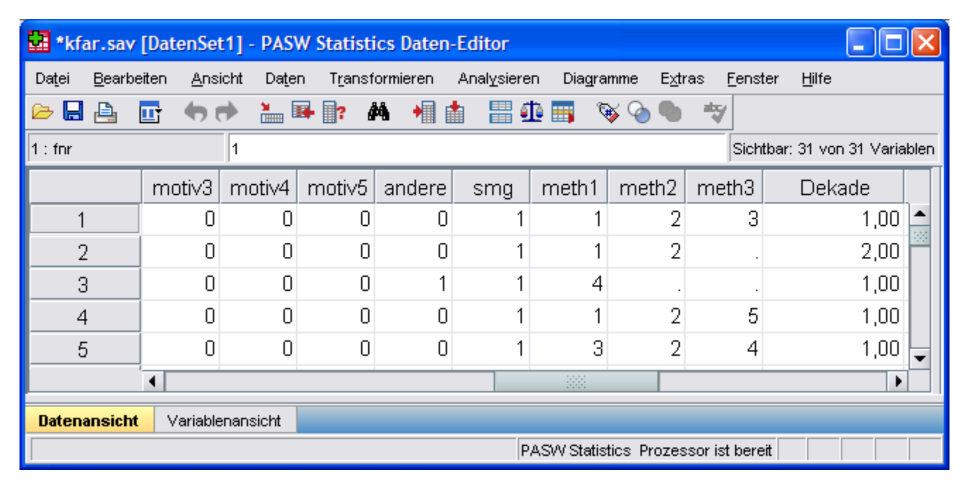


Damit ist die Rekodierung vollständig spezifiziert. Quittieren Sie die Subdialogbox mit **Weiter**. Da wir das KFA-Transformationsprogramm sukzessive aufbauen wollen, müssen Sie nun in der Dialogbox **Umkodieren in andere Variablen** auf den Schalter **Einfügen** klicken, um die implizit definierten Kommandos zu produzieren. Wir erhalten ein Syntaxfenster mit folgendem Inhalt:

```
DATASET ACTIVATE DatenSet1.
RECODE gebj (1960 thru 1969=1) (1970 thru 1979=2) INTO Dekade.
EXECUTE.
```

Das erste Kommando macht das **DatenSet1** zur Arbeitsdatei und soll verhindern, dass die nachfolgenden Kommandos auf ein ungeeignetes Datenset treffen. Hinter das RECODE-Kommando, das die eigentliche Umkodierung bewirkt, hat SPSS noch ein EXECUTE gesetzt, dessen Rolle in Abschnitt 6.3 erläutert wird.

Unabhängig von den guten Argumenten für das Transformationsprogramm gibt es in Ihrer aktuellen Lernphase einen Grund, die obige **Umkodieren**-Dialogbox per **OK**-Schalter zu quittieren oder die zugehörigen Kommandos jetzt schon ausführen zu lassen: Sie können den Effekt auf die Arbeitsdatei sofort beobachten, statt bis zum Abschicken des kompletten Transformationsprogramms warten zu müssen. Weil keine Konflikte mit unserer langfristigen Strategie zu befürchten sind, kehren wir (z.B. über den Symbolschalter) zur **Umkodieren**-Dialogbox zurück und quittieren sie mit **OK**. Anschließend befindet sich am rechten Rand der Arbeitsdatei die neue Variable DEKADE:



6.2.2 Technische Details

Obwohl das Umkodieren eine simple Datentransformation ist, sind bei der praktischen Anwendung doch einige technische Details zu beachten:

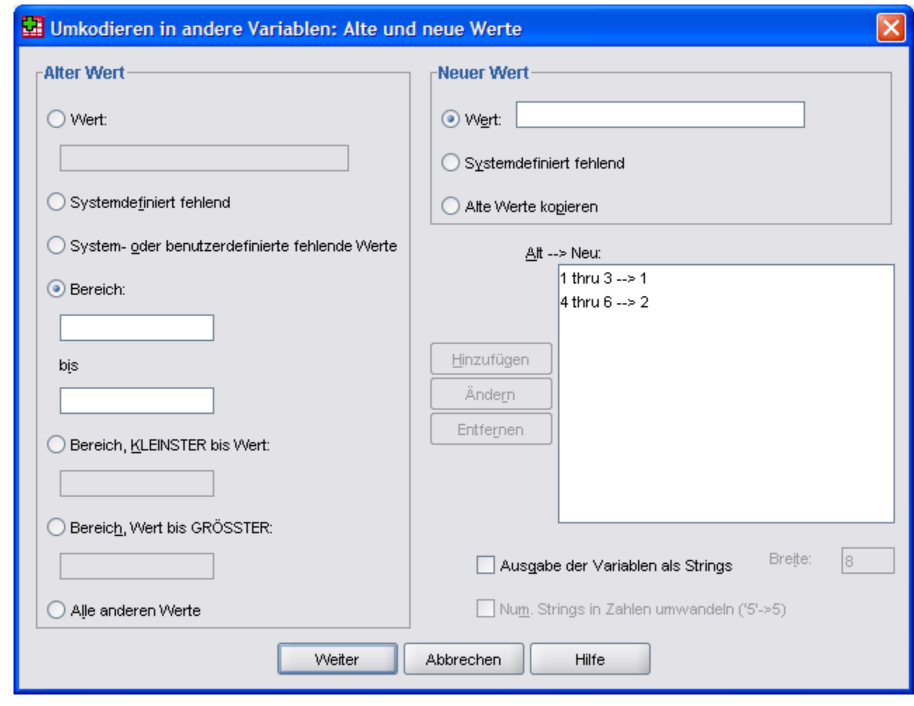
- Man kann bei einem Einsatz der Dialogbox **Umkodieren in andere Variablen** beliebig viele Variablen gleichzeitig umkodieren.
- Bei der Spezifikation der alten Werte, die auf einen neuen Wert abgebildet werden sollen, kann man angeben:
  - Einen einzelnen **Wert**
  - **Systemdefiniert fehlend** (SYSMIS)
    So ist es also möglich, den systemseitigen MD-Indikator auf einen anderen Wert umzusetzen.
  - **System- oder benutzerdefinierte fehlende Werte**
    Alle MD-Indikatoren werden umgesetzt.
  - Den **Bereich** von einem ersten Wert bis zu einem zweiten Wert (inklusive Grenzwerte)
    Bei allen Bereichen (auch den anschließend behandelten halboffenen Bereichen) ist zu beachten, dass im Bereich liegende benutzerdefinierte MD-Indikatoren einbezogen werden. Dies lässt sich z.B. mit einer einleitenden Ersetzungsvorschrift (MISSING = COPY) verhindern. Um diese Vorschrift per Dialogbox zu erzeugen, wählt man als alten Wert **System- oder benutzerdefinierte fehlende Werte** und als neuen Wert **Alte Werte kopieren** (siehe unten).
  - Den **Bereich** vom kleinsten Wert in der Stichprobe bis zu einem bestimmten Wert (inklusive Grenzwert)
  - Den **Bereich** von einem bestimmten Wert bis zum größten Wert in der Stichprobe (inklusive Grenzwert)

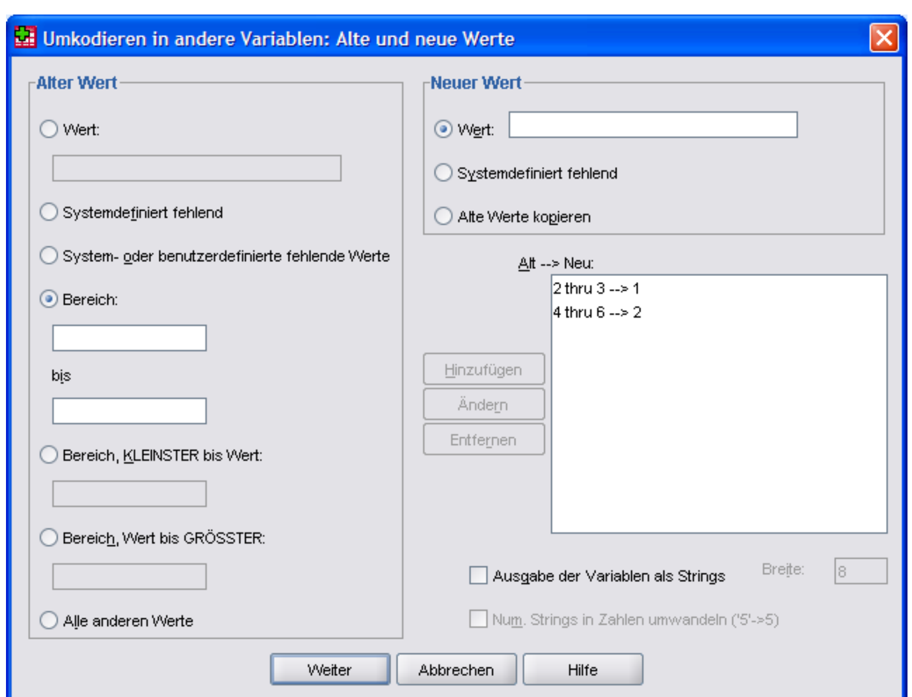
- **Alle anderen Werte**
  Damit sind alle in keiner anderen Ersetzungsvorschrift genannten Werte angesprochen (inklusive MD-Indikatoren, auch SYMIS). Um zu verhindern, dass auch MD-Indikatoren einbezogen werden, muss man diese Werte zuvor in einer speziellen Ersetzungsvorschrift behandeln, z.B. (MISSING = COPY). **Alle anderen Werte** kann nur in *einer* Ersetzungsvorschrift angegeben werden. Diese wird von SPSS in der Liste aller Ersetzungsvorschriften automatisch an die letzte Stelle gesetzt und damit bei der Ausführung zuletzt abgearbeitet.

- Als neuen Wert, auf den die alten Werte einer Ersetzungsvorschrift abgebildet werden sollen, können Sie angeben:
  - Einen **Wert**
  - **Systemdefiniert fehlend** (SYSMIS)
    Dann werden alle zugehörigen alten Werte auf SYSMIS umgesetzt.
  - **Alte Werte kopieren**
    Diese Möglichkeit steht nur beim Umkodieren in *andere* Variablen zur Verfügung und bewirkt für die zugehörigen alten Werte eine unveränderte Übernahme. Dies ist besonders nützlich, wenn die alten Werte mit **Alle anderen Werte** spezifiziert worden sind.
- Sie können beliebig viele Ersetzungsvorschriften festlegen. SPSS bringt diese automatisch in eine sinnvolle Ordnung.
- Wenn beim Umkodieren in andere Variablen eine *neue* Variable entsteht, so wird diese zunächst initialisiert, d.h. für alle Fälle wird in der neuen Spalte der Arbeitsdatei der Wert SYSMIS eingetragen (vgl. Abschnitt 6.1.3). Durch die *erste zutreffende* Ersetzungsregel wird bei einem Fall der Initialisierungswert durch den zugehörigen neuen Wert überschrieben. Alle weiteren (eventuell ebenfalls zutreffenden) Ersetzungsregeln werden bei diesem Fall ignoriert. Wird der alte Wert eines Falles in keiner Übersetzungsregel angesprochen, dann bleibt bei der neuen Variablen der Initialisierungswert SYSMIS stehen. Dies würde in obigem Beispiel etwa einem 1980 geborenen Untersuchungsteilnehmer passieren.
- Benutzerdefinierte MD-Indikatoren werden wie gültige Werte behandelt!
  Ist z.B. beim **Umkodieren in dieselben Variablen** für eine Variable der Wert 99 als benutzerdefinierter MD-Indikator deklariert, und wird die 99 rekodiert zur 98, dann **bleibt** die 99 ein MD-Indikator der Variablen, und die 98 wird **nicht** zum MD-Indikator. Eventuell muss also nach der Rekodierung die Variablendeklaration angepasst werden. Oben wurde schon erläutert, wie man bei Bereichen alter Werte die unerwünschte Mitbehandlung von benutzerdefinierte MD-Indikatoren verhindern kann.

6.2.3 Übungen

1) In den beiden folgenden Dialogboxen, die wir allerdings in unserem Projekt *nicht* ausführen wollen, wird jeweils eine Umkodierung der Fachbereichsvariablen (FB) in eine andere (neue) Variable spezifiziert. Hätten die beiden Dialogboxen denselben Effekt?





2) Bei unserem LOT-Fragebogen wurden die Fragen 3, 4, 5, und 12 aus messtechnischen Gründen umgepolt (negativ formuliert). Indem eine optimistische Antwort abwechselnd durch Zustimmung oder Ablehnung zum Ausdruck kommt, wird vermieden, dass systematische Ja- oder Neinsager einen extremen Optimismuswert erhalten. Bevor wir einen Mittelwert aus den LOT-Fragen als Optimismus-Schätzwert errechnen können, müssen die negativ gepolten Variablen folgendermaßen umkodiert werden:

$$5 \rightarrow 1$$
$$4 \rightarrow 2$$
$$2 \rightarrow 4$$
$$1 \rightarrow 5$$

Wählen Sie den Menübefehl:

Transformieren > Umkodieren in dieselben Variablen

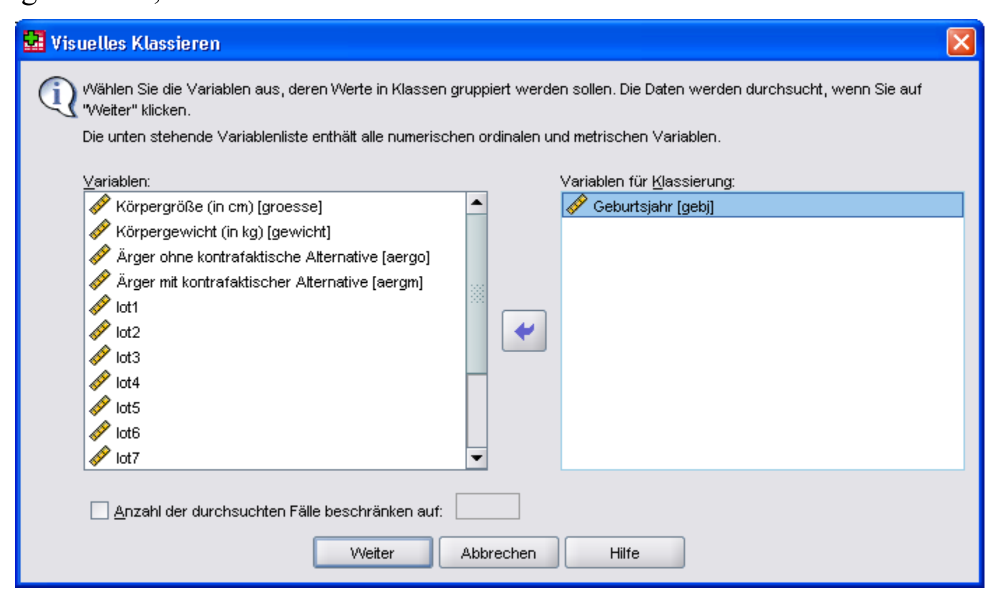
Quittieren Sie die bearbeitete Dialogbox **Umkodieren in dieselben Variablen** nicht mit **OK**, sondern mit **Einfügen**, damit das zugehörige RECODE-Kommando in das Syntaxfenster eingetragen wird, in dem wir gerade unser Transformationsprogramm aufbauen. Machen Sie sich klar, warum die Abbildungsvorschrift „3 → 3“ beim Umkodieren **In dieselben Variablen** überflüssig ist, beim Umkodieren in andere (neue) Variablen aber unbedingt erforderlich wäre.

6.2.4 Visuelles Klassieren

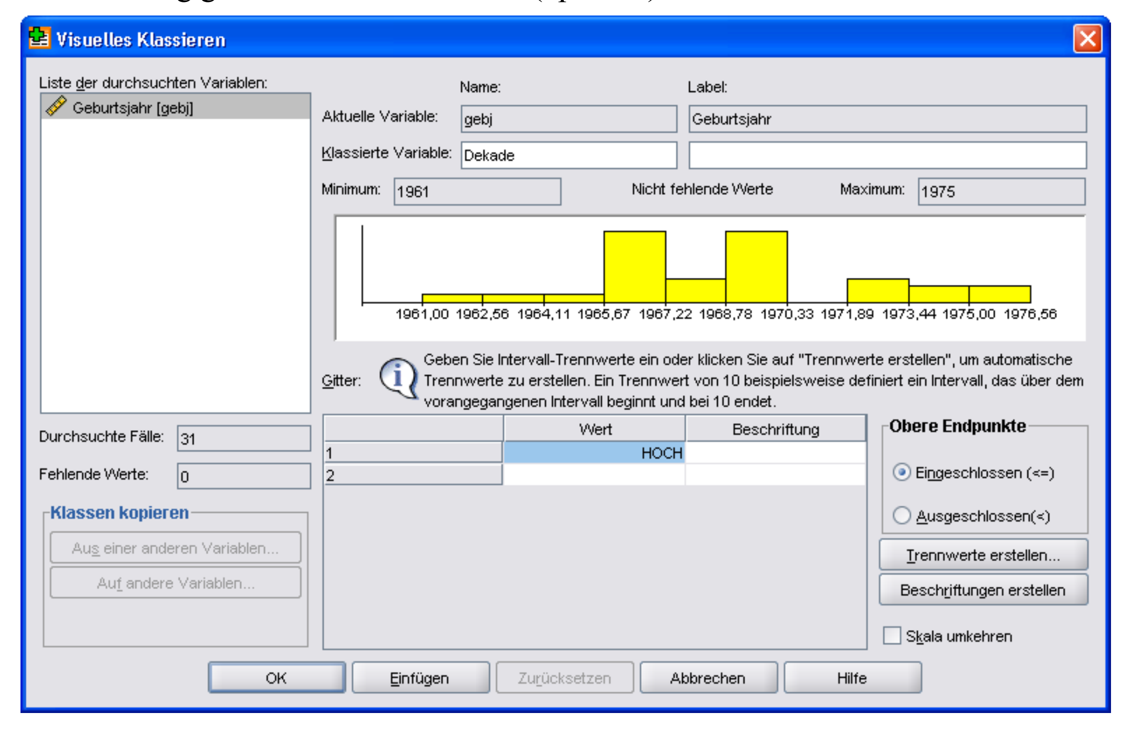
Über den Menübefehl

Transformieren > Visuelles Klassieren

ist ein Assistent zur Unterstützung der Klassenbildung zugänglich. Im ersten Schritt wählt man die Ausgangsvariable, z.B.:

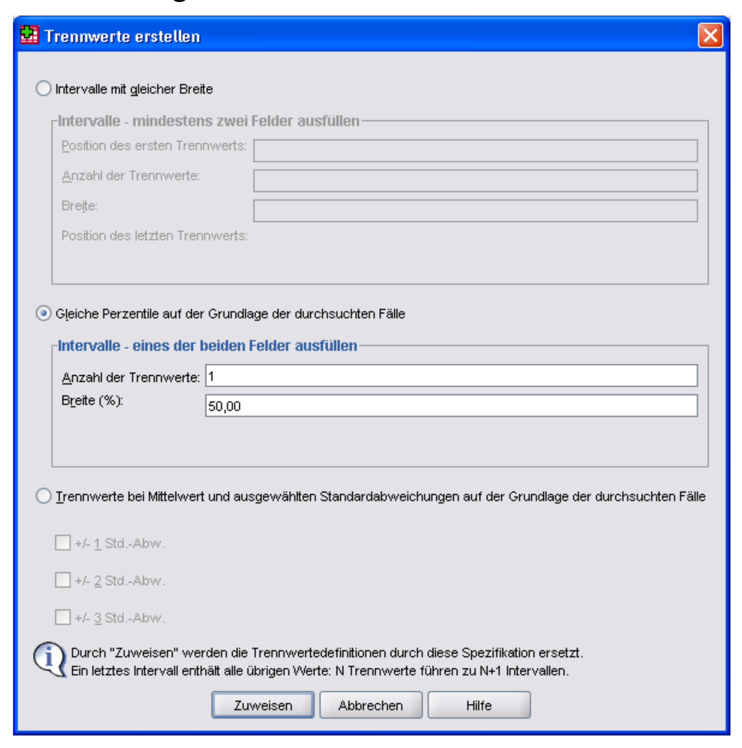


Im nächsten Dialog gibt man den Namen und (optional) das Label für die Zielvariable an:

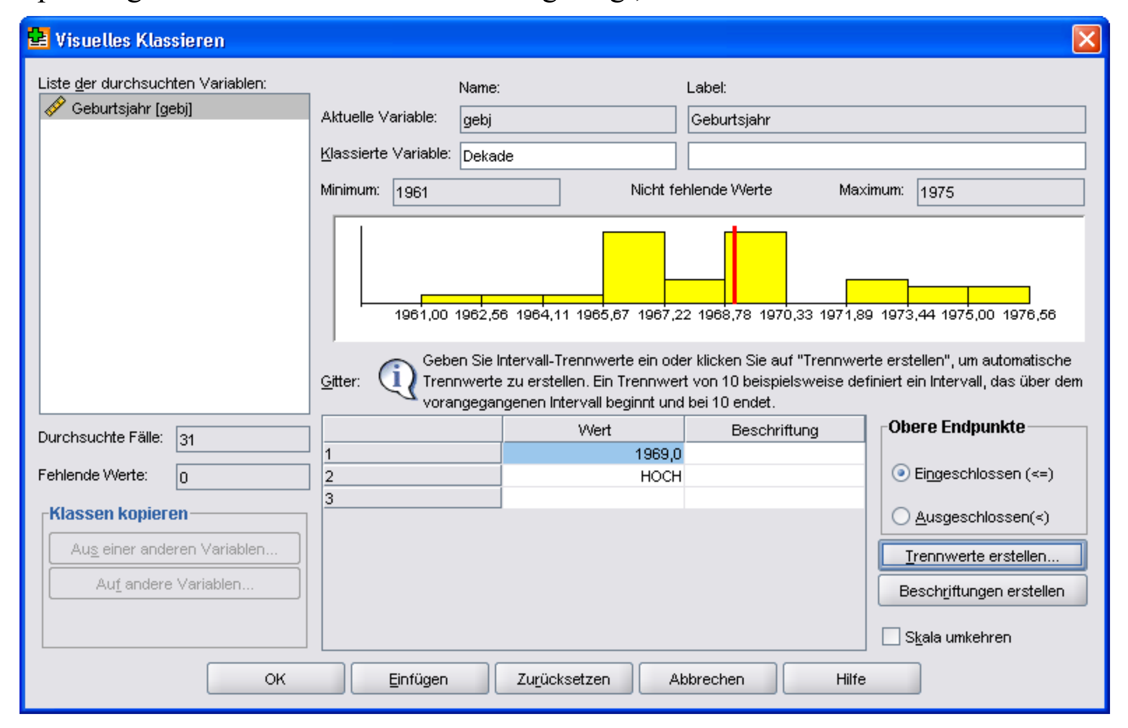


Ein Histogramm gibt eventuell Anregungen zur Aufteilung, und mit dem Kontrollkästchen **Skala umkehren** könnte man im Beispiel dafür sorgen, dass die Klasse mit den niedrigsten Geburtsjahren den höchsten Wert erhält.

Nach einem Klick auf den Schalter **Trennwerte erstellen** kann man im folgenden Dialog z.B. die Bildung von zwei annähernd gleich stark besetzten Klassen veranlassen:



Im Hauptdialog werden nun die Trennwerte angezeigt, z.B.:



Über den Schalter **Einfügen** erhält man u.a. das vom Assistenten erstellte RECODE-Kommando, z.B.:

```
RECODE  gebj (MISSING=COPY) (LO THRU 1969.0=1) (LO THRU HI=2) (ELSE=SYSMIS) INTO Dekade.
```

Es führt im Beispiel zum selben Ergebnis wie unsere eigene Variante (siehe Abschnitt 6.2.1) und demonstriert, wie man durch die geschickte Anordnung von Abbildungsvorschriften mit überlappenden Intervallen alter Werte dafür sorgt, dass *alle* alten Werte angesprochen werden.

6.3 Zur Rolle des EXECUTE-Kommandos

Wenn Sie eine **Umkodieren**-Dialogbox mit **OK** quittieren, dann führt SPSS per Voreinstellung die angeforderte Rekodierung sofort in der Arbeitsdatei aus. Obwohl dieses Verhalten sehr nahe liegend erscheint, gibt es eine erwägenswerte Alternative. Zum Rekodieren muss SPSS nämlich die Arbeitsdatei vollständig durchlaufen, was bei einer großen Stichprobe durchaus einige Zeit in Anspruch nehmen kann. Bei einer nächsten und übernächsten Transformationsanweisung (z.B. Rekodierung oder Neuberechnung) ist jeweils ein weiterer Durchlauf fällig. Dabei könnte SPSS zeitsparend *alle* Transformationen in einer *einzigen* Datenpassage erledigen. Diese könnte so lange aufgeschoben werden, bis durch die Anforderung einer Statistikprozedur das Durchackern der Daten unvermeidlich wäre. Genau in dem zuletzt beschriebenen, zeitökonomischen Sinn funktionieren seit jeher die SPSS-Transformationskommandos: Sie werden vorgemerkt und erst bei der nächsten Prozedur gemeinsam ausgeführt. Allerdings kann dieses zeitoptimierte Verhalten SPSS-Neulinge verwirren. Daher setzt die SPSS-Bedienoberfläche hinter jedes per Dialogbox produzierte Transformationskommando per Voreinstellung automatisch ein EXECUTE-Kommando, welches die *sofortige Ausführung* aller offenen Transformationen erzwingt. Wenn wir z.B. eine **Umkodieren**-Dialogbox mit **OK** quittieren, verarbeitet der SPSS-Prozessor im Hintergrund ein RECODE- *und* ein EXECUTE-Kommando. Das erste bewirkt nur eine Arbeitsvorbereitung, das zweite erzwingt die Ausführung der vorbereiteten Arbeit. Quittieren wir dieselbe Dialogbox mit **Einfügen**, erscheinen die beiden Kommandos im Syntaxfenster (siehe oben).[1]

Bei der in diesem Manuskript vorgestellten Arbeitsweise sind die von SPSS produzierten EXECUTE-Kommandos in der Regel überflüssig. Aufgrund der heutzutage verfügbaren Rechenleistung lohnt es sich allerdings nur bei einer sehr großen Arbeitsdatei, diese Kommandos aus einem automatisch produzierten Programm zu entfernen.

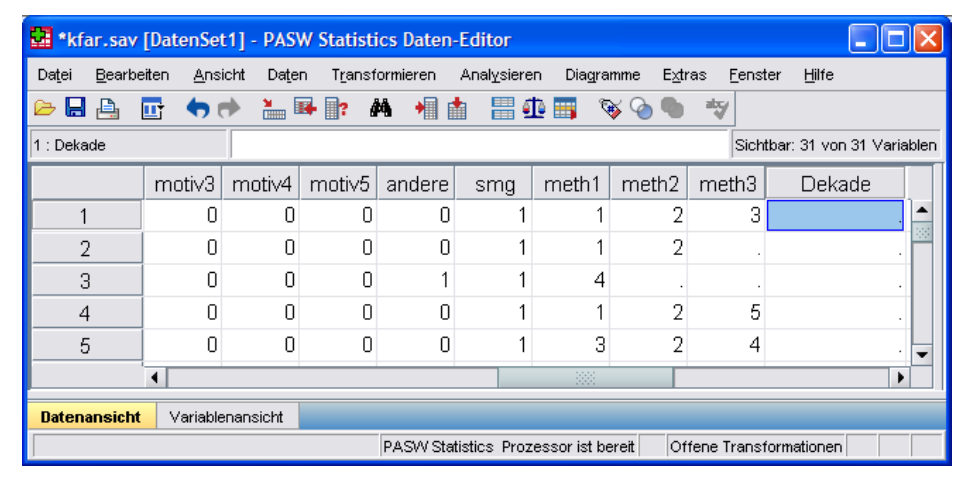
Beim Arbeiten mit dem Syntaxfenster kann es zu dem folgenden, recht frustrierenden Erlebnis kommen: Sie lassen wohlgeformte Transformationskommandos ausführen, doch im Datenfenster stellt sich nur ein partieller Erfolg ein. Zwar erscheinen die neu anzulegenden Variablen, doch haben alle Fälle den Wert SYSMIS, z.B.:

---

[1] Man kann nach

Bearbeiten > Optionen > Daten

im Rahmen **Optionen für Transformieren und Zusammenfügen** mit der Option **Werte vor Verwendung berechnen** die voreingestellte EXECUTE-Inflation abstellen. Dann zeigt SPSS das oben beschriebene zeitoptimierte Verhalten, führt also z.B. nach dem Quittieren einer **Umkodieren**-Dialogbox mit **OK** das zugrunde liegende RECODE-Kommando zunächst noch *nicht* aus, sondern reiht es in die Warteschlange der offenen Transformationen ein. Diese werden vom SPSS-Prozessor erst dann ausgeführt, wenn er ein Prozedur- oder ein EXECUTE-Kommando erhält.



Die Ursache ist dann meist: Sie haben nach den Transformationskommandos noch kein Prozedur- oder EXECUTE-Kommando ausführen lassen, so dass SPSS zwar die neue Variablen initialisiert, aber noch keine Werte ermittelt hat. In dieser Situation wird in der Statuszeile angezeigt, dass **Offene Transformationen** zur Bearbeitung anstehen. Sie können deren Ausführung erzwingen, indem Sie im Syntaxfenster ein EXECUTE-Kommando abschicken oder folgenden Menübefehl wählen:

Transformieren > Offene Transformationen ausführen

Es soll nicht verschwiegen werden, dass hier für SPSS-Neulinge Schwierigkeiten auftauchen, die bei rein Dialogbox-orientierter Arbeitsweise und voreingestelltem EXECUTE-Einsatz nicht entstehen können.

Für angehende SPSS-Profis soll noch erwähnt werden, dass EXECUTE-Kommandos *innerhalb eines Blocks von Transformationsanweisungen* durchaus bedeutsam sein können. In dem folgenden (manuell erstellten) Beispiel wird mit Hilfe des Transformationskommandos SELECT IF jeder zweite Fall aus der Arbeitsdatei entfernt:

```
compute nr = $casenum.
execute.
select if (mod(nr,2) = 1).
execute.
```

Lässt man jedoch das erste EXECUTE-Kommando weg, entfernt das Programm *alle* Fälle mit Ausnahme des ersten.

6.4 *Berechnung von Variablen nach mathematischen Formeln*

In der Dialogbox **Variable berechnen** bzw. im äquivalenten COMPUTE-Kommando wird ein numerischer Ausdruck (z.B. GROESSE - 100) definiert und einer Ergebnisvariablen zugewiesen. Dabei kann man eine *neue* Variable erzeugen oder eine vorhandene verändern.

6.4.1 Beispiel

Sie sollen später anhand unserer Stichprobe untersuchen, ob die Trierer Studierenden im Mittel wenigstens das folgende Idealgewicht auf die Waage bringen (Nullhypothese)

$$\textit{Gewicht (in kg)} \overset{!}{=} \textit{Größe (in cm) - 100}$$

oder ob sie relativ zu dieser Formel zu leicht sind (Alternativhypothese). Zur Prüfung dieser Frage mit einem t-Test für verbundene Stichproben muss die Arbeitsdatei um eine neue Variable, z.B. IDGEW genannt, erweitert werden, deren Werte nach der Formel

GROESSE - 100

aus der Körpergröße zu berechnen sind. Anschließend enthält die (Fälle × Variablen)-Datenmatrix in der Arbeitsdatei u.a. die beiden folgenden Variablen:

GROESSE	IDGEW
163	63
158	58
174	74
182	82
.	.
.	.
.	.
176	76
176	76
170	70
169	69

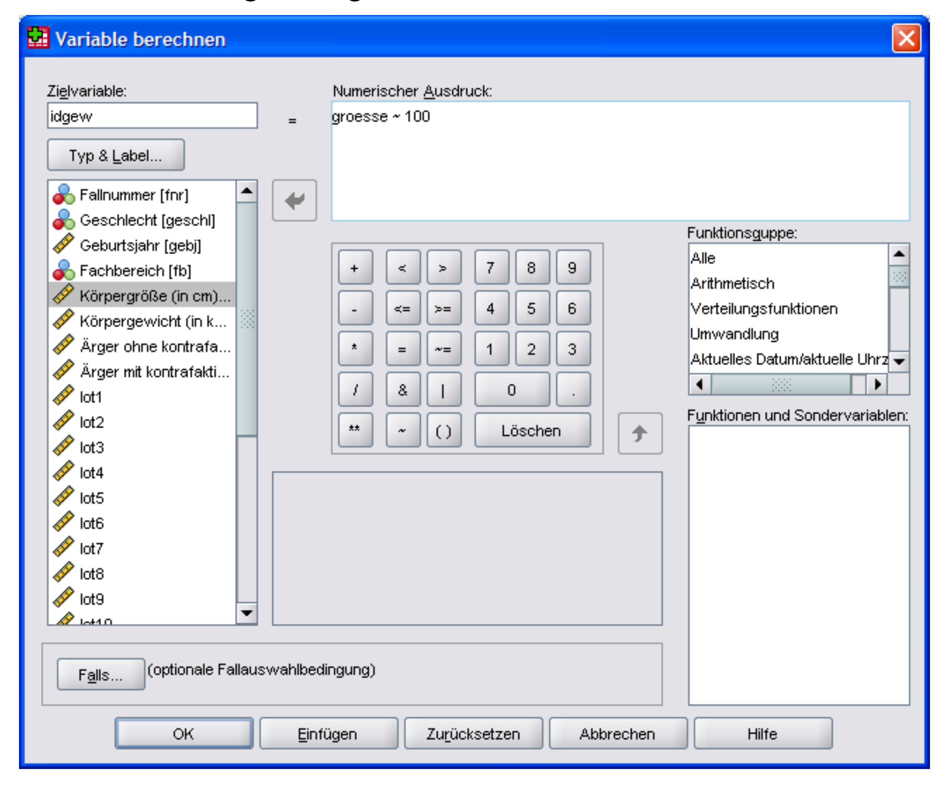
Starten Sie zum Definieren der neuen Variablen die Dialogbox **Variable berechnen** mit:

Transformieren > Variable berechnen

Tragen Sie zunächst im Feld **Zielvariable** den Namen für die neu in die Arbeitsdatei aufzunehmende Variable ein (IDGEW), und schreiben Sie dann in das Feld **Numerischer Ausdruck** die Definitionsvorschrift (GROESSE - 100), wobei einige Schreibhilfen zur Verfügung stehen:

- Der Variablenname kann aus einer Liste per Transportschalter, Drag & Drop oder Doppelklick übernommen werden.
- Mit Hilfe einer virtuellen Tastatur können Sie das Minuszeichen und die Zahl 100 auch per Maus eingeben.

Anschließend sollte Ihre Dialogbox ungefähr so aussehen:



Die Dialogbox bietet über unsere momentanen Bedürfnisse hinausgehend auch die in SPSS verfügbaren Funktionen (siehe unten) und spezielle Systemvariablen (z.B. **$Casenum** für die fort-

laufende Fallnummer in der Arbeitsdatei) in **Funktionsgruppen** geordnet zum Transport in das Feld **Numerischer Ausdruck** an, so dass man bei der Verwendung von Funktionen das Nachschlagen und Tippfehler vermeiden kann.

Rufen Sie nun mit dem gleichnamigen Schalter die Subdialogbox **Typ und Label** auf, und tragen Sie dort zur Variablen IDGEW die **Beschriftung** *Idealgewicht nach der Formel: Größe - 100* ein:



Quittieren Sie die Subdialogbox mit **Weiter** und die Hauptdialogbox mit **Einfügen**. Daraufhin erhalten Sie im Syntaxfenster ein COMPUTE - und ein VARIABLE LABELS - Kommando:

```
COMPUTE idgew = groesse - 100.
VARIABLE LABELS idgew 'Idealgewicht nach der Formel: Größe - 100'.
EXECUTE.
```

6.4.2 Technische Details

6.4.2.1 Numerischer Ausdruck

Im Texteingabefeld **Numerischer Ausdruck** der Dialogbox **Variable berechnen** sind wir trotz der SPSS-Scheibhilfen im Wesentlichen wieder in das „Syntaxzeitalter" zurückgeworfen: Auf der weißen Fläche ist ein sprachlicher Ausdruck nach gewissen Syntaxregeln zu formulieren. Zum Glück sind uns aber numerische Ausdrücke aus der Schule wohlbekannt.[1]

Ein numerischer Ausdruck im Sinne von SPSS darf folgende Bestandteile enthalten:

- bereits definierte Variablennamen
- Zahlen
- arithmetische Operatoren:
  - Addition (+)
  - Subtraktion (-)
  - Multiplikation (*)
  - Division (/)
  - Potenzfunktion (**)
- Klammern
- Funktionen

[1] Zwar gibt es gewisse Unterschiede zwischen mathematischen *Gleichungen* (z.B. $y = a + b \cdot x$) und EDV-sprachlichen *Zuweisungen* (z.B. compute x = x + 2.), doch sind die Regeln für die numerischen Ausdrücke auf den *rechten* Seiten weitgehend identisch.

6.4.2.1.1 Numerische Funktionen

In numerischen Ausdrücken können Sie zahlreiche Funktionen verwenden, die numerische Variablen oder Zahlen als Argumente (in den folgenden Syntaxdarstellungen vertreten durch den Platzhalter *arg*) verarbeiten.[1] Diese Funktionen lassen sich in mehrere Gruppen einteilen, aus denen jeweils einige wichtige Vertreter genannt werden sollen:

- **Arithmetische Funktionen**, z.B.:
  - ABS(*arg*) — Absoluter Wert
  - EXP(*arg*) — Exponentialfunktion
  - LG10(*arg*) — Dekadischer Logarithmus
  - LN(*arg*) — Natürlicher Logarithmus
  - MOD(*arg1*,*arg2*) — Rest aus der Division von *arg1* durch *arg2*
  - RND(*arg*) — Auf die nächst gelegene ganze Zahl gerundeter Wert
  - SQRT(*arg*) — Quadratwurzel

  Beispiel:
  ```
  compute logi = exp(3+1.2*x)/(1+exp(3+1.2*x)).
  ```
  Hier wird eine spezielle logistische Funktion der Variablen X definiert.

- **Statistische Funktionen**, z.B.:
  - MEAN[.*n*](*arg1*, *arg2*[, ...]) — Arithmetisches Mittel
  - MAX[.*n*](*arg1*, *arg2*[, ...]) — Maximum
  - MIN[.*n*](*arg1*, *arg2*[, ...]) — Minimum
  - SD[.*n*](*arg1*, *arg2*[, ...]) — Standardabweichung
  - SUM[.*n*](*arg1*, *arg2*[, ...]) — Summe

  Regeln:
  - Die eckigen Klammern schließen optionale Angaben ein.
  - Der Funktionsparameter *n* hat folgende Bedeutung: Wenn bei einem Fall mindestens *n* valide Argumente vorliegen, wird der Funktionswert berechnet. Ansonsten wird dem Fall der Wert SYSMIS zugewiesen. Wird *n* nicht angegeben, gilt die sehr liberale Voreinstellung Eins.
    Zwei häufige Fehler beim Einsatz des Minimalanforderungsparameters *n*:
    - Punkt zwischen dem Funktionsnamen und *n* vergessen
      Dieser Funktionsaufruf:
      ```
      mean45(sport to angeln)
      ```
      hat denselben Effekt wie der Aufruf
      ```
      mean(sport to angeln)
      ```
    - Leerzeichen zwischen dem Funktionsnamen und dem Punkt gesetzt
      Dieser Funktionsaufruf:
      ```
      mean .45(sport to angeln)
      ```
      führt zu einer Fehlermeldung.
  - Mit „[, ...]“ wird zum Ausdruck gebracht, dass die Liste der Argumente optional beliebig verlängert werden darf.
  - Sie können eine Serie von Variablen, *die in der Arbeitsdatei hintereinander stehen*, bequem auf folgende Weise in einer Argumentenliste angeben:

    erste TO *letzte*

    Es kommt nicht auf die alphanumerische Sortierung der Variablennamen an, sondern auf die Reihenfolge der Variablen in der Arbeitsdatei.

[1] SPSS kennt auch zahlreiche Funktionen für String- und Datums-Variablen, die aber aus Zeitgründen in diesem Kurs nicht behandelt werden. Informieren Sie sich bei Bedarf im Hilfesystem, z.B. über eine Indexsuche nach dem Stichwort *Funktionen*.

Beispiel: `compute mfrei = mean.45(sport to angeln).`

Wenn für einen Fall bei den Variablen SPORT bis ANGELN, die in der Arbeitsdatei hintereinander stehen, mindestens 45 valide Argumente vorliegen, wird deren Mittelwert der Variablen MFREI zugewiesen, ansonsten wird der MD-Indikator SYSMIS zugewiesen.

Beachten Sie den wesentlichen Unterschied zwischen den gerade beschriebenen statistischen *Funktionen* und den Statistik*prozeduren*, mit denen wir z.B. die Verteilungsanalysen durchgeführt haben:

- Wenn wir in der Dialogbox **Häufigkeiten** (erreichbar über **Analysieren > Deskriptive Statistiken > Häufigkeiten**) z.B. den Mittelwert der Variablen GEWICHT anfordern, werden die (validen) Gewichtsangaben aller Fälle in der Stichprobe gemittelt. Es werden also die Ausprägungen *einer Variablen* über *alle Fälle* gemittelt. SPSS arbeitet sich *senkrecht* durch eine komplette Variable bzw. Spalte der Arbeitsdatei. Es resultiert ein einziger Stichprobenkennwert, welcher im Ausgabefenster erscheint.
- Mit der statistischen Funktion MEAN können wir für *jede einzelne Person* z.B. den Mittelwert über *mehrere LOT-Variablen* berechnen lassen. SPSS geht *waagerecht* vor, wobei dasselbe Verfahren *auf jeden Fall, d.h. auf jede Zeile* der Datenmatrix angewendet wird. Die statistische Funktion MEAN erzeugt (oder modifiziert) eine Variable, d.h. eine Spalte in der Arbeitsdatei, in die für jeden Fall sein eigenes Berechnungsergebnis eingetragen wird.

- **Funktionen für fehlende Werte**, z.B.:

- NMISS(*arg1*[, ...])	Anzahl fehlender Werte bei den aufgelisteten Variablen
- VALUE(*arg*)	Es wird der Wert der Variablen *arg* geliefert, wobei *benutzerdefinierte* MD-Deklarationen ignoriert werden.

Regeln:
- Mit „[, ...]“ wird zum Ausdruck gebracht, dass die Liste der zu untersuchenden Variablen optional beliebig verlängert werden darf.
- Mit dem Schlüsselwort TO kann bequem eine Serie von Variablen angegeben werden (siehe obige Erläuterungen bei den statistischen Funktionen).

Beispiel: - `compute nmfrei = nmiss(sport to angeln).`

Der numerische Ausdruck liefert die Anzahl der fehlenden Werte (SYMIS oder benutzerdefiniert) bei den Variablen SPORT bis ANGELN, die in der Arbeitsdatei hintereinander stehen.

- **Pseudozufallszahlengeneratoren**, z.B.:

- NORMAL(*arg*)	Die Funktion liefert normalverteilte Zufallszahlen mit Mittelwert Null und Standardabweichung *arg*.
- UNIFORM(*arg*)	Die Funktion liefert gleichverteilte Zufallszahlen im Intervall von Null bis *arg*.

Beispiel:

```
COMPUTE av = NORMAL(1).
EXECUTE.
T-TEST
  GROUPS=geschl(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=av
  /CRITERIA=CIN(.95).
```

Die Kommandos in diesem Beispiel wurden mit Hilfe von Dialogboxen erzeugt (Schalter **Einfügen**). Im COMPUTE-Kommando wird die standardnormalverteilte Zufallsvariable AV erstellt. Es ist klar, dass Frauen und Männer bei AV denselben Erwartungswert (Populationsmittelwert) Null haben. Damit können wir ausprobieren, wie sich der t-Test für unabhängige Stichproben bei Gültigkeit der Nullhypothese identischer Erwartungswerte verhält. Die Dialogbox zu diesem t-Test erhält man mit **Analysieren > Mittelwerte vergleichen > T-Test bei unabhängigen Stichproben**.

Wenn Ihnen die Erläuterungen zu diesem Beispiel „spanisch“ vorkommen, hilft Ihnen vielleicht der Abschnitt 7.1 weiter, wo einige Grundprinzipien der Inferenzstatistik erläutert werden. Mit Gruppenvergleichen beschäftigen wir uns „offiziell“ in Abschnitt 8.

Hinweis: Bei NORMAL und UNIFORM wird ein Pseudozufallszahlengenerator verwendet, der per Voreinstellung mit dem festen Wert 2000000 startet und damit stets dieselben Zahlen liefert. Ein alternativer Startwert, der andere Zufallszahlen zur Folge hat, kann so gewählt werden:

- mit dem Menübefehl:
  Transformieren > Zufallszahlengeneratoren
- oder mit dem SPSS-Kommando:
  SET SEED=*n*.

6.4.2.1.2 Regeln für die Bildung numerischer Ausdrücke

Auch bei Verwendung der Dialogbox **Variable berechnen** müssen wir die numerischen Ausdrücke im Wesentlichen selbst formulieren. Dabei sind folgende Regeln zu beachten:

- Die **Auswertungsreihenfolge** hängt von der Priorität der Operatoren ab. Es gilt folgende Rangordnung:

  Priorität 1: Funktionen
  Priorität 2: Potenzieren (**)
  Priorität 3: Multiplikation (*), Division (/) und Vorzeichen-Minus (z.B.: "-b")
  Priorität 4: Addition (+), Subtraktion (-)

  Bei gleicher Priorität erfolgt die Auswertung von links nach rechts. Eine alternative Auswertungsreihenfolge kann durch Klammern erzwungen werden: Klammerausdrücke werden zuerst ausgewertet. Bei geschachtelten Klammern erfolgt die Auswertung von innen nach außen.

- Bei Funktionen mit mehreren Argumenten müssen die einzelnen Argumente **durch jeweils genau ein Komma** (optional ergänzt durch Leerzeichen) getrennt werden.

  Beispiel: `compute mabc = mean.2(a,b, c).`

- Obwohl SPSS *im Daten- und im Ausgabefenster* das ländertypische Dezimaltrennzeichen benutzt, bei uns also das Komma, müssen in numerischen Ausdrücken gebrochene Zahlen generell mit Dezimal**punkt** geschrieben werden:

  Richtig: `2.75`
  Falsch: `2,75`

  Dies gilt sowohl für das Feld **Numerischer Ausdruck** der Dialogbox **Variable berechnen** als auch für das COMPUTE-Kommando in einem Syntaxfenster. Es kann also durchaus passieren, dass Sie ein und dieselbe gebrochene Zahl im Datenfenster (als Wert eines Falles für eine bestimmte Variable) mit Dezimal*komma* und in der Dialogbox **Variable berechnen** (z.B. als Konstante in einer Berechnungsanweisung) mit Dezimal*punkt* schreiben müssen.

- Bei den meisten Funktionen sind auch numerische Ausdrücke als Argumente zugelassen. Beispiel: `compute albmax = max(a, ln(b)).`

6.4.2.2 *Sonstige Hinweise*

SYSMIS als Ergebnis eines numerischen Ausdrucks

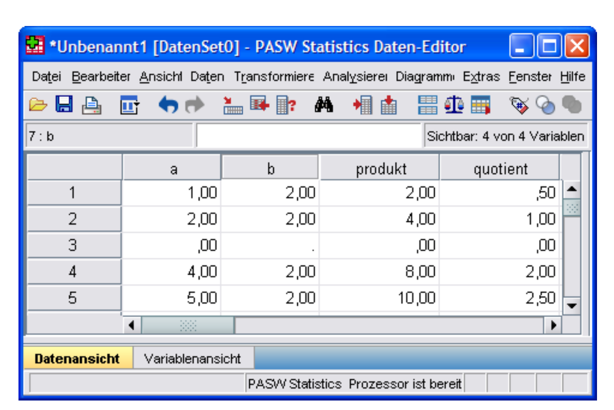
Durch eine Berechnungsanweisung wird der Wert des numerischen Ausdrucks auch dann der Zielvariablen zugewiesen, wenn dieser Wert gleich SYSMIS ist (z.B. bei fehlenden Argumenten). Dieses Vorgehen ist kompatibel mit dem in Abschnitt 6.1.3 beschriebenen Initialisierungsprinzip für *neue* numerische Variablen. Ist die Zielvariable bereits *vorhanden*, bleibt bei missglückter Berechnung des numerischen Ausdrucks keinesfalls der alte Wert bestehen, sondern es wird sinnvollerweise SYSMIS zugewiesen.

Rechnen mit fehlenden Werten

Fehlt bei einem Fall zur Berechnung eines numerischen Ausdrucks eine Argumentvariable, dann erhält die Ergebnisvariable den Wert SYSMIS. Ausnahmen sind die folgenden SPSS-eigenen Regeln für das „Rechnen“ mit fehlenden Werten:

- 0 * unbekannt = 0
  Diese Regel ist schlau, denn für beliebige reelle Zahlen x gilt:
  $$0 \cdot x = 0$$
- 0 / unbekannt = 0
  Diese Regel ist kritisierbar, denn:
  $$\frac{0}{x} = \begin{cases} 0 & \text{für } x \neq 0 \\ \text{undefiniert} & \text{für } x = 0 \end{cases}$$

Im folgenden Datenfenster hat der dritte Fall (mit dem Wert 0 bei der Variablen A und einem fehlenden B-Wert) für das Produkt A * B und den Quotienten A / B von SPSS den Ergebniswert Null erhalten:



	a	b	produkt	quotient
1	1,00	2,00	2,00	,50
2	2,00	2,00	4,00	1,00
3	,00	.	,00	,00
4	4,00	2,00	8,00	2,00
5	5,00	2,00	10,00	2,50

6.4.3 Übungen

1) Welche Werte haben die folgenden numerischen Ausdrücke?

(3 + 4) / 2
3 + 4 / 2
(3**2 / 2) + 4
3**2 / 2 + 4

2) Erstellen Sie im KFA-Projekt die Variablen, auf die sich unsere zentralen Hypothesen beziehen (vgl. Abschnitt 1.3):

   - Berechnen Sie die Variable LOT als arithmetisches Mittel der (nötigenfalls rekodierten!) LOT-Variablen 1, 3, 4, 5, 8, 9, 11 und 12. Die restlichen Fragen dienen nicht zur Messung von Optimismus, sondern sollen verhindern, dass der Zweck des Fragebogens deutlich wird. Dies könnte das Antwortverhalten verzerren.[1] Tolerieren Sie bei der Berechnung des Mittelwertes bis zu zwei fehlende Werte.
   - Berechnen Sie die Variable AERGAM als arithmetisches Mittel der beiden Ärgervariablen und die Variable AERGZ als Ärgerzuwachs auf Grund der kontrafaktischen Alternative. AERGAM benötigen wir zum Testen der differentialpsychologischen Hypothese. Beim geplanten Test der allgemeinpsychologischen Hypothese wird letztlich mit einem Einstichproben-t-Test geprüft, ob der Erwartungswert (Populationsmittelwert) der Variablen AERGZ signifikant größer als Null ist. Man kann den Test zwar bequem mit der SPSS-Prozedur zum t-Test für verbundene Stichproben durchführen, ohne die Variable AERGZ explizit berechnen zu müssen, doch bietet diese Prozedur keine Möglichkeit, die Normalverteilungsvoraussetzung des Tests (vgl. Abschnitt 7.1) zu prüfen. Daher berechnen wir AERGZ explizit und prüfen die Verteilungsvoraussetzung mit der Prozedur zur explorativen Datenanalyse (siehe Abschnitt 7.3).

   Rufen Sie jeweils mit dem Menübefehl:

   Transformieren > Variable berechnen

   die zuständige Dialogbox auf. Quittieren Sie Ihre Eintragungen nicht mit **OK**, sondern mit **Einfügen**, damit die zugehörigen COMPUTE-Kommandos in das Syntaxfenster eingetragen werden, in dem gerade das Transformationsprogramm zum KFA-Projekt entsteht.

   Weil SPSS eine Folge von mehreren Kommandos stets in der natürlichen Reihenfolge abarbeitet, wird beim späteren Ablauf unseres Transformationsprogramms z.B. die für einige Items angeordnete Rekodierung bereits erledigt sein, wenn das COMPUTE-Kommando zur LOT-Berechnung ausgeführt wird.

3) Erstellen Sie eine Variable namens BMI mit dem aus Körpergröße und Körpergewicht nach folgender Formel

$$\frac{Gewicht\ (in\ kg)}{Größe^2\ (in\ m)}$$

   berechneten **Body Mass Index**. Wir werden später im Rahmen unserer ernährungsphysiologischen Begleitstudien (siehe Abschnitt 8) der Frage nachgehen, ob beim BMI Geschlechtsunterschiede bestehen.

3) Berechnen Sie aus dem Geburtsjahr der Untersuchungsteilnehmer das Alter, wobei zu berücksichtigen ist, dass die Manuskriptstichprobe aus dem Jahr 2000 stammt. Wir haben bei der Datenerhebung nach dem Geburtsjahr gefragt, weil manche Auskunftspersonen diese Information leichter und genauer liefern können als das Alter. Bei der Forschungsarbeit und in Ergebnisberichten ist das Alter jedoch handlicher. Außerdem ist zu befürchten, dass mit dem Wissen um den Erhebungszeitpunkt irgendwann das Wissen um das Alter der Befragten verloren geht.

[1] Die von Scheier & Carver (1985) verwendete Verschleierungstechnik kann sicher in speziellen Fällen zur Verbesserung der Datenqualität beitragen, soll aber hier keinesfalls als Routinetechnik empfohlen werden.

6.5 Bedingte Datentransformation

Gelegentlich ist es erforderlich, eine Datenmodifikation auf diejenigen Fälle zu beschränken, die eine bestimmte Bedingung erfüllen. Wir benötigen z.B. im KFA-Projekt eine solche Möglichkeit, um bei den Motivations- und Methodenvariablen das bisher vertagte Problem der fehlenden Werte adäquat behandeln zu können (siehe Abschnitt 1.4.3.2).

Manchmal ist es angebracht, für mehrere disjunkte Teilmengen der Gesamtstichprobe jeweils spezifische Transformationen durchzuführen (Fallunterscheidung). Z.B. könnte man im Rahmen einer Untersuchung zum Essverhalten bei der Berechnung einer Idealgewichts aus der Körpergröße bei Frauen und Männern unterschiedliche Formeln anwenden.

In den SPSS - Transformations-Dialogboxen erreichen Sie über den Schalter **Falls** eine Subdialogbox zur Definition einer Bedingung, unter der die Transformation ausgeführt werden soll. Sie können z.B. eine bedingte Umkodierung (vgl. Abschnitt 6.2), Berechnung (vgl. Abschnitt 6.4) oder Werteauszählung (vgl. Abschnitt 6.6) vornehmen.

Wenn unter ein und derselben Bedingung gleich *mehrere* Transformationen vorgenommen werden sollen, muss diese Bedingung in allen benötigten Transformations-Dialogboxen, wiederholt werden. Ähnlich umständlich ist die Realisation von Fallunterscheidungen mit Hilfe der Transformations-Dialogboxen. Für solche Aufgaben bietet die SPSS-Kommandosprache mit der DO IF - ELSE IF - END IF - Kontrollstruktur bessere Lösungen. Diese lassen sich jedoch nicht komplett mit Dialogboxen generieren.

6.5.1 Beispiel

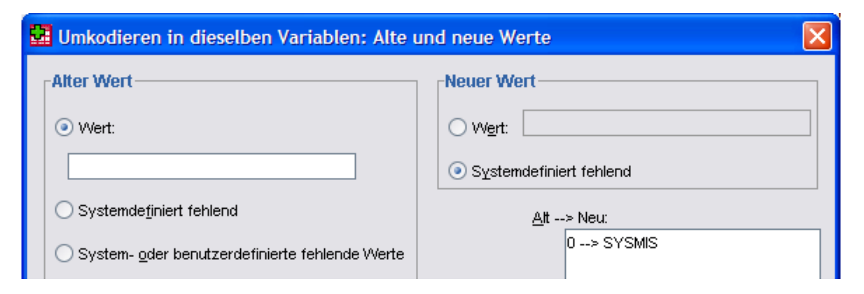
In diesem Abschnitt soll endlich das MD-Problem bei den Motivationsvariablen gelöst werden. Wir haben bei den Variablen MOTIV1 bis MOTIV5 und ANDERE die angekreuzten Kästchen mit Eins und die leeren Kästchen mit Null kodiert. Ein Fall mit Nullen bei MOTIV1 bis MOTIV5 *und* ANDERE hat aber offenbar den Fragebogenteil 4a komplett ausgelassen, denn:

- In der Stichprobe befinden sich ausschließlich Kursteilnehmer.
- Aufgrund der Restkategorie (Variable ANDERE) sind alle möglichen Motive zur Kursteilnahme berücksichtigt.
- Menschen führen ein aufwändiges Verhalten nur aus, wenn sie einen Grund dafür haben.

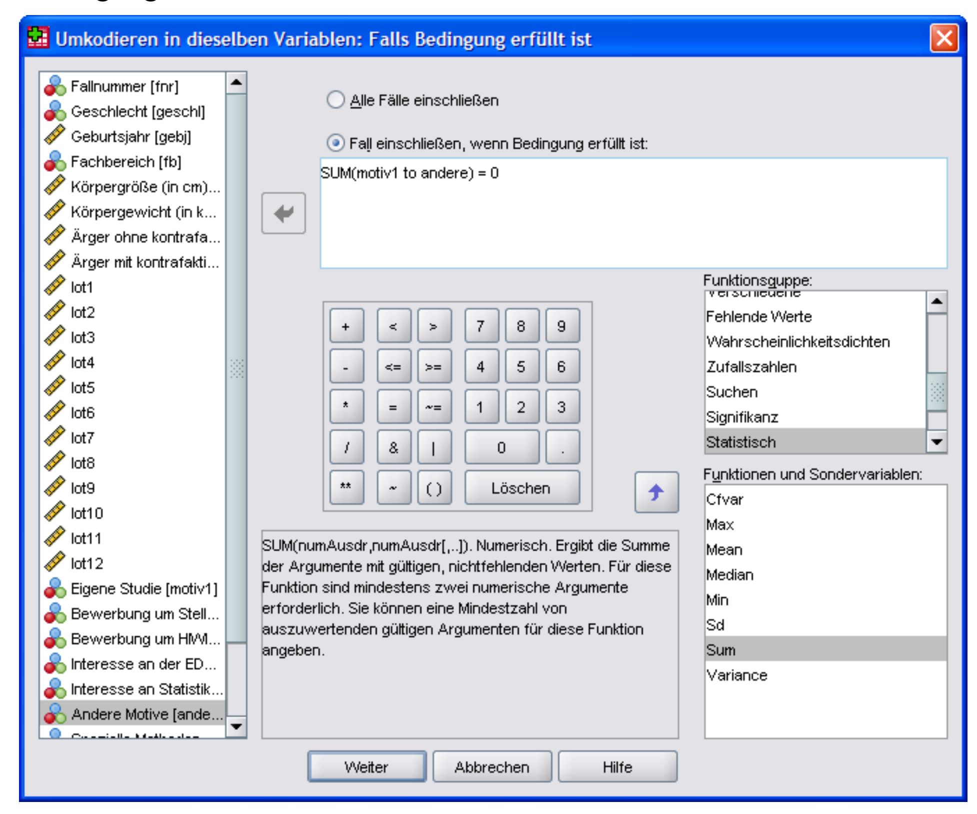
Daher sollten für genau diese Fälle die Nullen bei den Variablen MOTIV1 bis MOTIV5 und ANDERE in SYSMIS umkodiert werden. Gehen Sie folgendermaßen vor:

- Wählen Sie den Menübefehl:

  Transformieren > Umkodieren in dieselben Variablen
- Transportieren Sie die Variablennamen MOTIV1 bis MOTIV5 und ANDERE in die Teilnehmerliste der **Umkodieren**-Dialogbox.
- Legen Sie in der Subdialogbox **Alte und neue Werte** die benötigte Abbildungsvorschrift fest:



- Öffnen Sie die **Falls**-Subdialogbox, markieren Sie die Option **Fall einschließen, wenn Bedingung erfüllt ist**, und tragen Sie in das darunter liegende Textfeld eine geeignete Bedingung ein, z.B.:



Aufgrund unserer Datenüberprüfung (siehe Abschnitt 4) können wir uns darauf verlassen, dass bei den Variablen MOTIV1 bis MOTIV5 und KEINE ausschließlich die Werte Null und Eins vorliegen. Daher ist die Summe dieser Variablen genau dann gleich Null, wenn jede einzelne Variable gleich Null ist.

Die obige Eintragung im Bedingungsfeld kann „semiautomatisch" z.B. folgendermaßen erzeugt werden:

- Wählen Sie die Funktionsgruppe **Statistisch**, markieren Sie die Funktion **Sum**, und klicken Sie auf den zugehörigen Transportschalter, so dass im Bedingungsfeld eine Vorlage für einen SUM-Funktionsaufruf erscheint:

  SUM(?,?)

- Transportieren Sie aus der Variablenliste MOTIV1 in das Bedingungsfeld, wobei in der Vorlage das markierte Fragezeichen automatisch durch den Variablennamen ersetzt wird.
- Ersetzen Sie das Komma in der Vorlage durch das Schlüsselwort TO, und komplettieren Sie die Liste durch den Variablennamen ANDERE, den Sie wiederum am besten aus der Liste in das Bedingungsfeld transportieren.
- Komplettieren Sie den Funktionsaufruf zu einer Bedingung, indem Sie ein Gleichheitszeichen und den Wert Null anhängen.

- Machen Sie **Weiter**, und quittieren Sie die Hauptdialogbox mit **Einfügen**.

Daraufhin wird Ihr Transformationsprogramm um die folgende Sequenz erweitert:

```
DO IF (SUM(motiv1 to andere) = 0).
RECODE motiv1 motiv2 motiv3 motiv4 motiv5 andere (0=SYSMIS).
END IF.
EXECUTE.
```

Wenn Sie diese Kommandos ausführen lassen, gleichgültig ob direkt per **OK** in der **Umkodieren**-Dialogbox oder indirekt via Syntaxfenster, passiert bei jedem einzelnen Fall in der Stichprobe folgendes:

- SPSS prüft die Bedingung, die wir auch als **logischen Ausdruck** bezeichnen wollen.
- Ist bei einem Fall die Bedingung erfüllt, dann wird umkodiert, anderenfalls passiert nichts.

Weil die Variablen MOTIV1 bis MOTIV5 und ANDERE vor der Rekodierung garantiert nur Nullen oder Einsen als Werte aufweisen, hat unser logischer Ausdruck die Eigenschaft, in jedem Fall entweder wahr oder falsch zu sein. Das erscheint nach dem aussagenlogischen Axiom vom ausgeschlossenen Dritten als selbstverständlich, ist es aber in der empirischen Forschung z.B. wegen des nahezu allgegenwärtigen Problems fehlender Werte keineswegs. Generell kann z.B. der logische Ausdruck „GESCHL = 1“ folgende Wahrheitswerte annehmen:

- wahr ⇔ Der GESCHL-Wert ist gleich Eins.
- falsch ⇔ Der GESCHL-Wert ist eine von Eins verschiedene Zahl.
- unbestimmt ⇔ Der GESCHL-Wert fehlt.

Komplexere logische Ausdrücke (z.B. „LN(ML)/ANZ > 1“) können auch wegen undefinierter Funktionswerte unbestimmt sein (bei ML ≤ 0 oder ANZ = 0).

Wenn Sie eine bedingte Transformationsanweisung verwenden, sollten Sie beachten, wie SPSS auf bestimmte und unbestimmte logische Ausdrücke reagiert:

- Ist der logische Ausdruck **wahr**, dann wird die Transformation ausgeführt.
  Im Fall einer bedingten Berechnung (COMPUTE-Kommando) wird der Ergebnisvariablen also der Wert des numerischen Ausdrucks zugewiesen. Die Zuweisung erfolgt auch dann, wenn der numerische Ausdruck den Wert SYSMIS hat.
- Ist der logische Ausdruck **falsch oder unbestimmt**, dann passiert **nichts**, d.h.:
  - Eine bereits vorhandene Ergebnisvariable behält für den betroffenen Fall ihren bisherigen Wert.
  - Bei einer neu definierten Variablen behält der betroffene Fall den Initialisierungswert SYSMIS.

6.5.2 Bedingungen formulieren

Der in obigem Beispiel aufgetretene logische Ausdruck war recht einfach aufgebaut, weil er nur aus einem einzigen Vergleich bestand. Obwohl Ihnen auch komplexere Exemplare (z.B. aus der Schule) wohlvertraut sein dürften, soll der Begriff *logischer Ausdruck* zur Klärung einiger Spezialprobleme etwas genauer beschrieben werden. Zunächst wird der einfachere Begriff *Vergleich* eingeführt.

6.5.2.1 Vergleich

Ein Vergleich besteht aus zwei numerischen Ausdrücken und einem Vergleichsoperator:

```
numerischer_ausdruck   vergleichsoperator   numerischer_ausdruck
```

Die bekannten **Vergleichsoperatoren** können in SPSS alternativ durch EDV-Varianten der mathematischen Symbole oder durch Schlüsselwörter dargestellt werden:

Symbol	Schlüsselwort	Bedeutung
=	EQ	gleich
<>	NE	ungleich
<	LT	kleiner als
<=	LE	kleiner oder gleich
>	GT	größer als
>=	GE	größer oder gleich

Beispiele:
```
beruf >= 4
beruf ge 4
```

6.5.2.2 Logischer Ausdruck

Aus dem einfachen Begriff *Vergleich* wird nun durch eine rekursive Definition der komplexere Begriff *logischer Ausdruck* konstruiert:

i) Jeder Vergleich ist ein logischer Ausdruck.

ii) Durch Anwendung des logischen Operators **NOT** auf einen logischen Ausdruck oder durch Anwendung der logischen Operatoren **AND** bzw. **OR** auf zwei logische Ausdrücke entsteht ein neuer logischer Ausdruck:

NOT *logischer_ausdruck*

logischer_ausdruck_1 AND *logischer_ausdruck_2*

logischer_ausdruck_1 OR *logischer_ausdruck_2*

Den Wahrheitswert eines zusammengesetzten logischen Ausdrucks erhält man aus den Wahrheitswerten der Argumente nach den Regeln für logische Operatoren, die in den so genannten Wahrheitstafeln festgelegt sind (siehe unten).

Es lassen sich sukzessiv beliebig komplexe logische Ausdrücke aufbauen, die für jeden konkreten Fall die Wahrheitswerte *wahr*, *falsch* oder *unbestimmt* haben können.

Beispiel:
```
(beruf >= 4)  and  (schule <> 7)
```

Mit unbestimmten Wahrheitswerten in logischen Ausdrücken verfährt SPSS analog zum Rechnen mit fehlenden Werten in numerischen Ausdrücken (siehe Abschnitt 6.4.2.2). Die folgenden Wahrheitstafeln sind gegenüber der klassischen Aussagenlogik um den Wahrheitswert *unbestimmt* erweitert (*la1* und *la2* seien logische Ausdrücke):

la1	NOT *la1*
wahr	falsch
falsch	wahr
unbestimmt	unbestimmt

la1	*la2*	*la1* AND *la2*	*la1* OR *la2*
wahr	wahr	wahr	wahr
wahr	falsch	falsch	wahr
wahr	unbestimmt	unbestimmt	wahr
falsch	wahr	falsch	wahr
falsch	falsch	falsch	falsch
falsch	unbestimmt	falsch	unbestimmt
unbestimmt	wahr	unbestimmt	wahr
unbestimmt	falsch	falsch	unbestimmt
unbestimmt	unbestimmt	unbestimmt	unbestimmt

6.5.2.3 Regeln für die Auswertung logischer Ausdrücke

Bei der Auswertung von logischen Ausdrücken gelten in SPSS folgende Regeln:

- Die Abarbeitungsreihenfolge hängt von der Priorität der Operatoren ab. Es gilt folgende Rangordnung:

  Priorität 1: Funktionen
  Priorität 2: Potenzieren (**)
  Priorität 3: Multiplikation (*), Division (/), Vorzeichen-Minus (z.B. -a)
  Priorität 4: Addition (+), Subtraktion (-)
  Priorität 5: Vergleichsoperatoren
  Priorität 6: NOT
  Priorität 7: AND
  Priorität 8: OR

- Bei gleicher Priorität: Abarbeitung von links nach rechts.
- Eine andere Auswertungsreihenfolge kann durch Klammern erzwungen werden.

Das obige Beispiel für einen zusammengesetzten logischen Ausdruck kann wegen der voreingestellten Abarbeitungsreihenfolge auch kürzer geschrieben werden:

```
beruf >= 4  and  schule <> 7
```

Die aus Computer-Sicht überflüssigen Klammern verbessern allerdings die Lesbarkeit des Ausdrucks für Menschen und reduzieren so das Fehlerrisiko.

6.5.3 Übung

Bei den Variablen METH1 bis METH3 haben wir zur Vereinfachung der Erfassung im Kodierplan festgelegt, dass „unbenutzte“ Variablen einfach leer bleiben sollen. Nun wollen wir aber bei Fällen mit regulärem Antwortmuster die SYSMIS - Werte durch Nullen ersetzen. Die Null soll z.B. bei der Variablen METH2 bedeuten: Die Option, einen zweiten Methodenwunsch zu äußern, wurde nicht genutzt.

Die folgende Tabelle, die wir in Abschnitt 1.4.3.2.4 vereinbart haben, legt im Einzelnen fest, was unter den möglichen Bedingungskonstellationen geschehen soll:

		Mindestens eine speziell interessierende Methode angegeben?	
		Ja	Nein
SMG	1	METH1 ... METH3: SYSMIS → 0 Bem.: Korrektes Antwortverhalten. Variablen zu nicht benutzten Optionen (gem. Kodierplan bisher auf SYSMIS) werden auf 0 gesetzt.	SMG: 1 → SYSMIS Bem.: Irreguläres Antwortverhalten. METH1 bis METH3 behalten SYMIS. SMG wird ebenfalls auf SYMIS gesetzt.
	0	SMG: 0 → 1 METH1 ... METH3: SYSMIS → 0 Bem.: Leicht irreguläres Antwortverhalten. Wir sind großzügig und setzen SMG auf 1.	METH1 ... METH3: SYSMIS → 0 Bem.: Korrektes Antwortverhalten. Die Variablen zu allen Optionen (gem. Kodierplan bisher auf SYSMIS) werden auf 0 gesetzt.
	SYSMIS	SMG: SYSMIS → 1 METH1 ... METH3: SYSMIS → 0 Bem.: Leicht irreguläres Antwortverhalten. Wir sind großzügig und setzen SMG auf 1 sowie die Variablen zu nicht benutzten Optionen auf 0.	Bem.: Irreguläres Antwortverhalten. Alle Variablen behalten den Wert SYSMIS.

In den beiden obersten Zeilen jeder Zelle sind die erforderlichen Korrekturen bei SMG bzw. METH1 bis METH3 angegeben. Erweitern Sie Ihr Programm **kfat.sps** um passende Transformationsanweisungen.

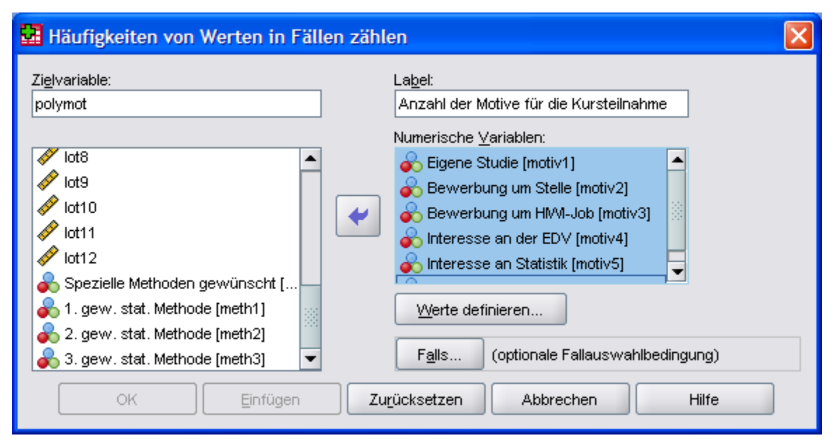
6.6 Häufigkeit bestimmter Werte bei einem Fall ermitteln

Mit dem Befehl **Werte in Fällen zählen** aus dem Menü **Transformieren** bzw. mit dem zugrunde liegenden COUNT-Kommando kann man eine Variable berechnen lassen, die für jeden Fall festhält, wie viele Variablen aus einer Liste mit *k* Elementen einen kritischen Wert haben. Die Definition der kritischen Werte erfolgt über eine ein- oder mehrelementige Werteliste. Das minimale Zählergebnis ist Null (keine Variable hat einen der kritischen Werte), und das maximale Ergebnis ist *k* (jede Variable hat einen kritischen Wert).

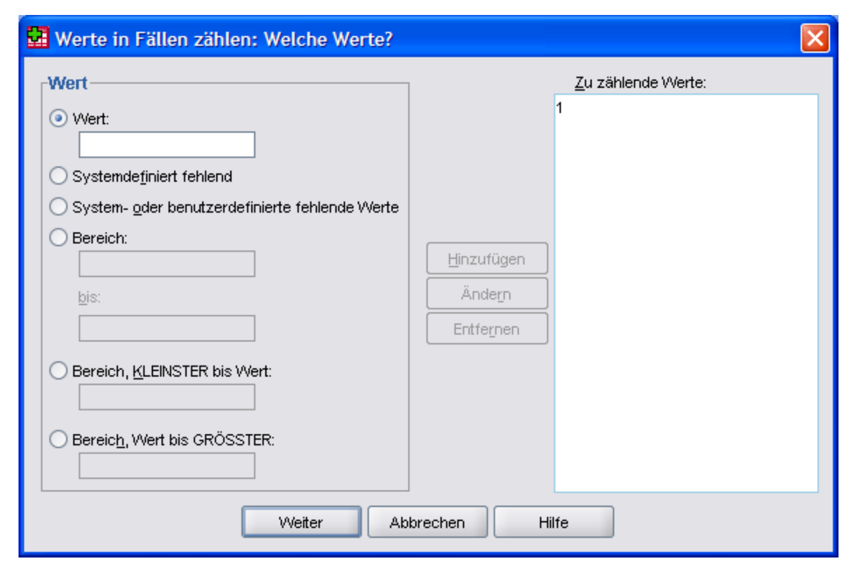
Wir wollen eine neue Variable namens POLYMOT berechnen lassen, die für jede Person festhält, wie viele Motive zur Kursteilnahme sie im Fragebogenteil 4a angegeben hat. Aktivieren Sie die Dialogbox **Häufigkeiten von Werten in Fällen zählen** mit

Transformieren > Werte in Fällen zählen

Vergeben Sie für die Zielvariable den Namen POLYMOT sowie das Label *Anzahl der Motive für die Kursteilnahme*, und transportieren Sie die Variablen MOTIV1 bis ANDERE in die Teilnehmerliste. Danach müsste Ihre Dialogbox ungefähr so aussehen:



Wechseln Sie jetzt mit dem Schalter **Werte definieren** in die Subdialogbox **Werte in Fällen zählen: Welche Werte?**, tragen Sie dort den kritischen **Wert** Eins ein, und klicken Sie auf **Hinzufügen**:



Die in dieser Subdialogbox angebotenen sonstigen Möglichkeiten zur Festlegung der Trefferwerte kennen wir übrigens schon aus der Subdialogbox **Umkodieren: Alte und neue Werte** (siehe Abschnitt 6.2).

Da SPSS eine Folge von mehreren Kommandos stets in der natürlichen Reihenfolge abarbeitet, wird beim späteren Ablauf unseres Transformationsprogramms die MD-Problematik bei den Variablen MOTIV1 bis ANDERE bereits gelöst sein, wenn die **Zählen**-Anweisung an die Reihe kommt. Bei Personen, die den Fragebogenteil 4a *nicht* korrekt bearbeitet haben, wird also gelten:

MOTIV1 = MOTIV2 = ... = ANDERE = SYSMIS

Wir müssen die folgende wichtige Eigenschaft der **Zählen**-Anweisung beachten: Ihre Ergebnisvariable hat *stets* einen validen Wert größer oder gleich Null. Wenn ein Fall z.B. bei allen kritischen Variablen den - nicht zu zählenden - Wert SYSMIS hat, resultiert das valide Ergebnis Null! In dieser Situation wissen wir aber *nichts* von den Motiven der Person und dürfen ihr keine Motivationslosigkeit (POLYMOT = 0) unterstellen.

Weil im konkreten Beispiel das Zählergebnis Null generell als irregulär einzustufen ist, könnten wir durch ein gewöhnliches (unbedingtes) Umkodieren

$0 \rightarrow$ SYSMIS

dafür sorgen, dass ein Fall bei POLYMOT den Wert SYSMIS erhält, wenn er den Fragebogenteil 4a nicht korrekt bearbeitet hat. Im Allgemeinen kann das Zählergebnis Null jedoch auch auf reguläre Weise zustande kommen, und auch ein von Null verschiedenes Zählergebnis kann MD-belastet sein. Daher ist es meist erforderlich, durch eine bedingte Datentransformation MD-belastete Zählergebnisse zu verhindern. Wir wollen das generelle Verfahren der Übung halber auch im aktuellen Beispiel einsetzen und formulieren mit Hilfe der in Abschnitt 6.4.2.1.1 beschriebenen Funktion NMISS die folgende Bedingung

NMISS(MOTIV1 TO ANDERE) = 0

Klicken Sie bitte in der Dialogbox **Häufigkeiten von Werten in Fällen zählen** auf den **Falls**-Schalter, und tragen Sie die vorgeschlagene Bedingung ein. Wenn Sie dann **Weiter** machen und die Hauptdialogbox mit **Einfügen** quittieren, erhalten Sie im Syntaxfenster die folgenden Kommandos:

```
DO IF  (NMISS(MOTIV1 TO ANDERE) = 0).
COUNT polymot=motiv1 motiv2 motiv3 motiv4 motiv5 andere(1).
VARIABLE LABELS  polymot 'Anzahl der Motive für die Kursteilnahme'.
END IF.
EXECUTE.
```

Was hier zählt, ist offenbar das COUNT-Kommando. Es enthält im Wesentlichen eine Liste der zu untersuchenden Variablen, gefolgt von einer eingeklammerten Liste der kritischen Werte. Das VARIABLE LABELS - Kommando hat SPSS aufgrund unserer Eintragung im **Label**-Textfeld erstellt. Das Zählergebnis wird nur dann ermittelt und der neuen Variablen POLYMOT als Wert zugewiesen, wenn die Bedingung im DO IF - Kommando erfüllt ist. Anderenfalls behält POLYMOT den Initialisierungswert SYSMIS.

6.7 Erstellung der Fertigdatendatei mit dem Transformationsprogramm

Aufgrund von nachvollzogenen Demonstrationsbeispielen bzw. Übungsaufgaben in den Abschnitten 6.2 (Erstellung von DEKADE durch Rekodierung von GEBJ, Umkodieren der negativ formulierten LOT-Fragen), 6.4 (Berechnung von IDGEW, LOT, AERGAM, AERGZ, BMI und ALTER), 6.5 (MD-Behandlung für die Motiv- und für die Methoden-Variablen) und 6.6 (Auszählen der Kursmotive) sollten jetzt alle vorläufig im KFA-Projekt benötigten Transformationskommandos in einem Syntaxfenster stehen.

6.7.1 Transformationsprogramm vervollständigen

Um daraus ein komfortables SPSS-Programm zu machen, das die Rohdatendatei **kfar.sav** selbständig einließt, die so entstandene Arbeitsdatei transformiert und schließlich als Fertigdatendatei **kfa.sav** auf die Festplatte schreibt, müssen wir an den Anfang des Syntaxfensters noch ein GET-Kommando zum Öffnen von **kfar.sav** und ans Ende noch ein SAVE-Kommando zum Sichern in **kfa.sav** setzen. Wie Sie das GET-Kommando produzieren können, haben Sie schon in Abschnitt 5.2 erfahren. Wenn Sie das Kommando jetzt erzeugen lassen, erscheint es am Ende des Syntaxfensters, und Sie müssen es an den Anfang verschieben.

Wir verzichten auf das automatisch erzeugte DATASET NAME - Kommando, streichen es also aus dem Programm. Ebenso löschen wir alle DATASET ACTIVATE - Kommandos aus dem Transformationsprogramm.

Um das SAVE-Kommando zu generieren, wechseln wir ins Datenfenster und aktivieren mit **Datei > Speichern unter** die zugehörige Dialogbox. Dann tragen wir den gewünschten Dateinamen **kfa.sav** ein und erzeugen mit **Einfügen** das benötigte SAVE-Kommando.

Hinweise zur Ausgabedatei eines Transformationsprogramms:

- Verwenden Sie niemals dieselbe Datei als Quelle und Ziel des Transformationsprogramms. Schreiben Sie also keinesfalls mit Ihrem Transformationsprogramm in die Rohdatendatei. Wenn Sie der Empfehlung in Abschnitt 6.1.2 folgend für die Rohdatendatei das Schreibschutzattribut gesetzt haben, kann dieses Desaster auch nicht versehentlich passieren.
- Bei der Ausführung des Transformationsprogramms darf für seine Ausgabedatei, also für die Fertigdatendatei, das Schreibschutzattribut natürlich nicht gesetzt sein.

Schließlich sollte Ihr Syntaxfenster ungefähr folgenden Inhalt haben:

```
GET
  FILE='U:\Eigene Dateien\SPSS\kfar.sav'.

* DEKADE.
RECODE gebj (1960 thru 1969=1) (1970 thru 1979=2) INTO Dekade.
EXECUTE.

* LOT-Fragen umkodieren.
RECODE lot3 lot4 lot5 lot12 (5=1) (4=2) (2=4) (1=5).
EXECUTE.

* IDGEW berechnen.
COMPUTE idgew = groesse - 100.
VARIABLE LABELS idgew 'Idealgewicht nach der Formel: Größe - 100'.
EXECUTE.

* LOT berechnen.
COMPUTE lot = MEAN.6(lot1,lot3,lot4,lot5,lot8,lot9,lot11,lot12).
VARIABLE LABELS lot 'LOT-Optimismus'.
EXECUTE.

* AERGAM berechnen.
COMPUTE aergam = (aergo + aergm)/2.
VARIABLE LABELS aergam 'Mittel der Ärger-Variablen'.
EXECUTE.

* AERGZ berechnen.
COMPUTE aergz =  aergm - aergo.
VARIABLE LABELS aergz 'Ärger-Zuwachs durch die KFA'.
EXECUTE.

* BMI berechnen.
COMPUTE bmi = gewicht / (groesse/100)**2.
VARIABLE LABELS bmi 'Body Mass Index'.
EXECUTE.

* Alter berechnen.
COMPUTE Alter = 2000 - gebj.
EXECUTE.

* MD-Behandlung für die Motiv-Variablen.
DO IF (SUM(motiv1 to andere) = 0).
RECODE motiv1 motiv2 motiv3 motiv4 motiv5 andere (0=SYSMIS).
END IF.
EXECUTE.

* MD-Behandlung für die Methoden-Variablen, Zelle (1,1) der Tabelle.
DO IF (smg=1 and nmiss(meth1 to meth3) < 3).
RECODE meth1 meth2 meth3 (SYSMIS=0).
END IF.
EXECUTE.

* MD-Behandlung für die Methoden-Variablen, Zelle (1,2) der Tabelle.
DO IF (smg=1 and nmiss(meth1 to meth3) = 3).
RECODE smg  (1=SYSMIS).
END IF.
EXECUTE.

* MD-Behandlung für die Methoden-Variablen, Zelle (2,1) der Tabelle.
DO IF ((smg = 0)  and  (nmiss(meth1 to meth3) < 3)).
RECODE smg  (0=1).
END IF.
EXECUTE .
DO IF ((smg = 0)  and  (nmiss(meth1 to meth3) < 3)).
RECODE meth1 meth2 meth3 (SYSMIS=0).
END IF.
EXECUTE.
```

```
* MD-Behandlung für die Methoden-Variablen, Zelle (2,2) der Tabelle.
DO IF (smg=0 and nmiss(meth1 to meth3) = 3).
RECODE meth1 meth2 meth3 (SYSMIS=0).
END IF.
EXECUTE.

* MD-Behandlung für die Methoden-Variablen, Zelle (3,1) der Tabelle.
DO IF ((nmiss(smg) = 1)  and  (nmiss(meth1 to meth3) < 3)).
RECODE smg  (SYSMIS=1).
END IF.
EXECUTE.
DO IF ((nmiss(smg) = 1)  and  (nmiss(meth1 to meth3) < 3)).
RECODE meth1 meth2 meth3 (SYSMIS=0).
END IF.
EXECUTE.

* POLYMOT berechnen.
DO IF (NMISS(motiv1 to andere) = 0).
COUNT polymot=motiv1 motiv2 motiv3 motiv4 motiv5 andere(1).
VARIABLE LABELS polymot 'Anzahl der Motive für die Kursteilnahme'.
END IF.
EXECUTE.

* Variablenattribute setzen.
formats dekade idgew aergz alter polymot (f8.0) aergam (f8.1) lot bmi (f8.2).
variable width dekade to polymot (7).
variable level dekade (ordinal) / idgew to polymot (scale).

SAVE OUTFILE='U:\Eigene Dateien\SPSS\kfa.sav'
   /COMPRESSED.
```

In diesen Lösungsvorschlag ist etwas Handarbeit eingeflossen:

- Zwischen manchen Kommandos sind der Übersichtlichkeit halber Leerzeilen eingefügt worden. Man darf aber auf keinen Fall *innerhalb* eines Kommandos (d.h. zwischen dem Kommandonamen und dem abschließenden Punkt) eine Leerzeile einfügen (vgl. Abschnitt 5.4).
- Die mit einem Sternchen (*) eingeleiteten Zeilen beinhalten *Kommentare*, die nachträglich eingefügt wurden, um die spätere Orientierung im Programm zu erleichtern (vgl. Abschnitt 5.4).

  Wichtig: Ein Kommentar hat ebenfalls Kommandostatus und muss daher unbedingt mit einem Punkt abgeschlossen werden. Sonst erstreckt sich der Kommentar bis zur nächsten Zeile, die entweder komplett leer ist oder mit einem Punkt endet.
- Eventuell legen Sie Wert darauf, dass auch die neu berechneten Variablen mit einer optimalen Anzahl von Dezimalstellen angezeigt werden. Eine manuelle Einstellung (vgl. Abschnitt 3.2.2) ist wenig attraktiv, weil unser Transformationsprogramm ja mit einiger Wahrscheinlichkeit mehrfach ausgeführt werden muss. Die bessere Alternative besteht darin, das Programm um ein FORMATS-Kommando zu erweitern, das die Attribute automatisch setzt:

  ```
  formats  dekade idgew aergz alter polymot (f8.0) aergam (f8.1) lot bmi (f8.2).
  ```

  Im Ausdruck „(f*b*.*d*)“ legt man mit *b* die Gesamtbreite der Wertausgabe (Attribut **Spaltenformat**) und mit *d* die Anzahl der Dezimalstellen fest. Weil bei numerischen Variablen die Gesamtbreite für uns irrelevant ist, haben wir bei den Rohvariablen auf eine Anpassung der Voreinstellung Acht verzichtet. So verfahren wir der Einheitlichkeit halber auch bei den abgeleiteten Variablen.

  Fügen Sie das FORMATS-Kommando am Ende des Transformationsprogramms ein (unmittelbar vor dem SAVE-Kommando).

- Mit den folgenden Kommandos wird die Breite der Datenfensterspalte (Attribut **Spalten**) und das Messniveau für die neuen Variablen eingestellt, wobei SCALE für Intervallskalenqualität steht:

  ```
  variable width  dekade to polymot (7).
  variable level  dekade (ordinal) / idgew to polymot (scale).
  ```

  Fügen Sie die Kommandos am Ende des Transformationsprogramms ein (unmittelbar vor dem SAVE-Kommando).

Beachten Sie bitte im Zusammenhang mit dem Thema Datensicherheit:

- Wenn zum Zeitpunkt der Programmausführung die Arbeitsdatei keinen Datenblatt - Namen hat, wird sie vom GET-Kommando ohne Nachfrage überschrieben!
- Das SAVE-Kommando überschreibt eine eventuell vorhandene Datei **kfa.sav** ohne Nachfrage, was jedoch bei der in diesem Manuskript vorgeschlagenen Arbeitsweise (vgl. Abschnitt 6.1.1) unproblematisch ist.

Damit ist das Transformationsprogramm zum KFA-Projekt fertig. Falls noch nicht geschehen, müssen Sie es unbedingt sichern, z.B. in das Verzeichnis **U:\Eigene Dateien\SPSS** unter dem oben vorgeschlagenen Dateinamen **kfat.sps**.

6.7.2 Transformationsprogramm ausführen

Lassen Sie das Transformationsprogramm ausführen, z.B. mit

Ausführen > Alles

Wenn alles glatt läuft, finden Sie anschließend im (aktiven) Ausgabefenster nur die protokollierten Kommandos und die folgende Warnung:

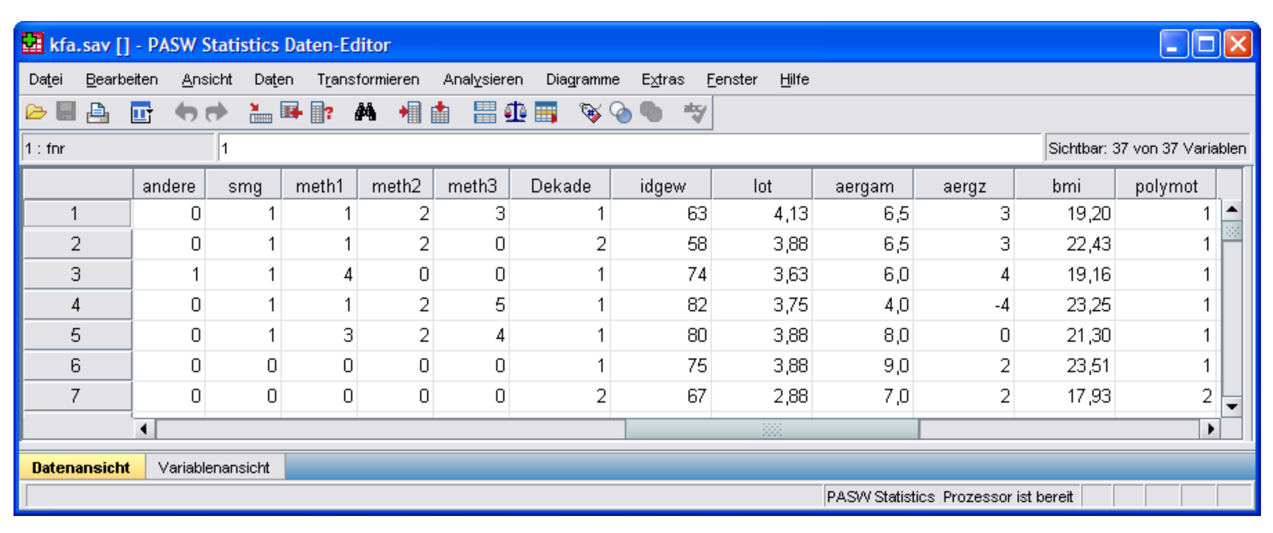
```
>Warnung Nr.  67.  Befehlsname: GET FILE
>Das Dokument wird bereits von einem anderen Benutzer oder Prozess verwendet.
>Wenn Sie das Dokument ändern, können so Änderungen anderer Benutzer
>überschrieben oder Ihre Änderungen durch andere überschrieben werden.
>Geöffnete Datei U:\Eigene Dateien\SPSS\kfar.sav
```

Diese Warnung bezieht sich auf das doppelte Öffnen der Rohdatendatei und ist irrelevant, weil das SAVE-Kommando am Ende des Transformationsprogramms das aktive Datenblatt mit der Fertigdatendatei verknüpft und somit die Verbindung mit der Rohdatendatei aufhebt.

Andere Warnungen oder Fehlermeldungen und/oder Warnungen müssen analysiert werden. Da alle Kommandos Ihres Programms von SPSS erstellt wurden, sollte Sie aber von dieser Mühe verschont bleiben.

Ältere Warnungen bzw. Fehlermeldungen sollten vor einem Lauf des Transformationsprogramms aus dem Ausgabefenster gelöscht werden, um Unklarheiten zu vermeiden.

Durch einen gelungenen Lauf unseres Transformationsprogramms entsteht ein unbenanntes Datenblatt, das mit der per SAVE erstellten Fertigdatendatei **kfa.sav** verbunden ist. Am rechten Rand der Datenmatrix sind die neuen Variablen zu finden, z.B.:



	andere	smg	meth1	meth2	meth3	Dekade	idgew	lot	aergam	aergz	bmi	polymot
1	0	1	1	2	3	1	63	4,13	6,5	3	19,20	1
2	0	1	1	2	0	2	58	3,88	6,5	3	22,43	1
3	1	1	4	0	0	1	74	3,63	6,0	4	19,16	1
4	0	1	1	2	5	1	82	3,75	4,0	-4	23,25	1
5	0	1	3	2	4	1	80	3,88	8,0	0	21,30	1
6	0	0	0	0	0	1	75	3,88	9,0	2	23,51	1
7	0	0	0	0	0	2	67	2,88	7,0	2	17,93	2

Das seit Beginn unserer Arbeit am Transformationsprogramm vorhandene, mit der Rohdatendatei verbundene Datenset ändert sich durch den Programmlauf *nicht*, weil es einen Namen erhalten hat (z.B. **DatenSet1**) und daher vom GET-Kommando des Transformationsprogramms nicht tangiert wird. Sie müssen also vor der Erfolgskontrolle das tatsächlich relevante Dateneditorfenster ansteuern (z.B. per **Fenster**-Menü).

Sie dürfen aber Ihre Erfolgskontrolle keinesfalls auf das Datenfenster beschränken, sondern müssen unbedingt das Ausgabefenster auf Fehlermeldungen und Warnungen überprüfen. SPSS stoppt nämlich die Programmausführung **nicht** beim Auftreten des ersten fehlerhaften Kommandos, sondern ignoriert das fehlerhafte Kommando und macht unverdrossen mit den nächsten Kommandos weiter. Diese arbeiten aber in der Regel aufgrund des vorangegangenen Fehlers mit falschen Zwischenergebnissen und produzieren Unsinn. Es kann also leicht passieren, dass nach einem fehlerbehafteten Lauf des Transformationsprogramms alle erwarteten neuen Variablen vorhanden sind, jedoch fehlerhafte Werte enthalten.

7 Prüfung der zentralen Projekt-Hypothesen

7.1 Entscheidungsregeln beim Hypothesentesten

In diesem Abschnitt werden einige Grundprinzipien der Inferenzstatistik am Beispiel unserer allgemeinpsychologischen Hypothese demonstriert. Dabei handelt es sich nicht um eine systematische Behandlung des Themas, die erheblich mehr Platz beanspruchen würde. Im Wesentlichen sollen die statistischen Entscheidungsregeln so präsentiert werden, dass sie mit Hilfe der SPSS-Ausgaben unmittelbar umgesetzt werden können. Zumindest in älteren Statistikbüchern findet man nämlich Formulierungen mit wenig Bezug zu den heute üblichen Ausgaben von Statistikprogrammen.

Wenn mit μ_O der Erwartungswert (Populationsmittelwert) des Merkmals AERGO und mit μ_M der Erwartungswert des Merkmals AERGM bezeichnet wird, dann lautet unser allgemeinpsychologisches KFA-Testproblem:

$$H_0: \mu_M \leq \mu_O \quad \text{vs.} \quad H_1: \mu_M > \mu_O$$

Mit Hilfe der Differenzvariablen AERGZ := AERGM – AERGO, deren Erwartungswert mit μ_Z bezeichnet werden soll, lässt sich das Testproblem äquivalent noch kompakter formulieren:

$$H_0: \mu_Z \leq 0 \quad \text{vs.} \quad H_1: \mu_Z > 0$$

Bei der Reformulierung wird die folgende Identität ausgenutzt (Linearität des Erwartungswerts):

$$\mu_Z = \mu_M - \mu_O$$

Wir setzen voraus, dass die Differenzvariable AERGZ normalverteilt ist mit dem Erwartungswert μ_Z und der Varianz σ_Z^2:

$$\text{AERGZ} \sim N(\mu_Z, \sigma_Z^2)$$

Für die *n* AERGZ-Beobachtungen in der Stichprobe nehmen wir an, dass sie durch **unabhängiges** „Ziehen“ aus der eben beschriebenen Population entstanden sind. Das schon in Abschnitt 1.1 betonte Unabhängigkeitsprinzip ist die zentrale Forderung in unserem **Stichprobenmodell** über die Gewinnung der empirischen Daten.

Bei der inferenzstatistischen Lösung des beschriebenen Testproblems verwendet man eine so genannte **Teststatistik *T*** (synonym: **Prüfgröße**), die aus den Stichprobendaten berechnet werden kann und folgende Eigenschaften besitzt:

- Sie ist indikativ für Abweichungen der wahren Populationsverteilung von der Nullhypothesenbehauptung und wächst tendenziell mit zunehmender Effektstärke

  $$dz := \frac{\mu}{\sigma}$$

  im Sinne der Alternativhypothese (vgl. Abschnitt 1.3.2). Sie quantifiziert also, wie gut bzw. schlecht die Nullhypothese mit den Stichprobendaten vereinbar ist.
- Es ist bekannt, welcher Verteilung die Teststatistik *T* bei *gültiger Nullhypothese* folgt, also bei $\mu_Z \leq 0$. Damit lässt sich für den konkreten Wert T_{emp} der Teststatistik in einer bestimmten Stichprobe berechnen, mit welcher Wahrscheinlichkeit eine Nullhypothesenpopulation Zufallsstichproben mit einer Teststatistikausprägung größer oder gleich T_{emp} liefert. Ist diese Wahrscheinlichkeit sehr klein, liegt der Schluss nahe, dass die konkret vorliegende Stichprobe *nicht* aus einer Nullhypothesenpopulation stammt.

In der oben beschriebenen Situation hat sich die folgende Teststatistik T_Z bewährt (mit Z als Abkürzung für AERGZ):

$$T_Z := \frac{\overline{Z}}{S_Z}\sqrt{n} \quad \text{mit} \quad \overline{Z} := \frac{1}{n}\sum_{i=1}^{n} Z_i \quad \text{und} \quad S_Z := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2}$$

Dabei ist $\overline{Z}$ das Stichprobenmittel und S_Z die Wurzel aus dem erwartungstreuen Schätzer der Populationsvarianz σ_Z^2.

Für das Stichprobenmittel $\overline{Z}$ (als Zufallsvariable aufgefasst) ergibt sich die Varianz

$$\text{Var}(\overline{Z}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} Z_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(Z_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Z^2 = \frac{n\sigma_Z^2}{n^2} = \frac{\sigma_Z^2}{n}$$

und damit die Streuung (der so genannte *Standardfehler*)

$$\sqrt{\text{Var}(\overline{Z})} = \frac{\sigma_Z}{\sqrt{n}}$$

Folglich schätzt $\frac{S_Z}{\sqrt{n}}$ den Standardfehler des Stichprobenmittelwerts, und T_Z ist gerade der Quotienten aus dem Stichprobenmittelwert und seinem geschätzten Standardfehler:

$$T_Z = \frac{\overline{Z}}{S_Z}\sqrt{n} = \frac{\overline{Z}}{\frac{S_Z}{\sqrt{n}}}$$

Teststatistiken von analoger Bauart sind uns schon bei den approximativen Tests zur Schiefe bzw. Wölbung einer Verteilung in Abschnitt 4.5 begegnet.

T_Z erfüllt die obigen Anforderungen:

- Die wesentliche, ergebnisabhängige T_Z –Komponente $\frac{\overline{Z}}{S_Z}$ ist offenbar ein Schätzer für die Effektstärke $\frac{\mu}{\sigma}$. Folglich wächst T_Z stochastisch mit zunehmender Effektstärke.
- Für $\mu_Z = 0$ besitzt T_Z (bei beliebigem Nebenparameter σ_Z^2) eine t-Verteilung mit n - 1 Freiheitsgraden. Damit kennen wir das Verhalten der Teststatistik *am Rand* der Nullhypothese. Dieses Wissen genügt, weil die bei einer Testentscheidung relevante Überschreitungswahrscheinlichkeit unter der H_0 (siehe unten) am Rand der Nullhypothese (also bei $\mu_Z = 0$) maximal wird. Ist sie am Rand klein genug, dann gilt dies auch für alle anderen Verteilungen in der Nullhypothese.

Aufgrund dieser Voraussetzungen kann man zu dem in einer konkreten Stichprobe erzielten Wert T_{emp} der Teststatistik T_Z die folgende **Überschreitungswahrscheinlichkeit** bestimmen:

> Mit welcher Wahrscheinlichkeit nimmt die Teststatistik T_Z bei einer Zufallsstichprobe der Größe n aus einer **Nullhypothesenpopulation** (genauer: bei $\mu_Z = 0$) einen Wert größer oder gleich T_{emp} an?

Diese Wahrscheinlichkeit wollen wir mit $\mathbf{P_{H_0}(T_Z \geq T_{emp})}$ bezeichnen. Sie wird von SPSS berechnet und in der Ausgabe zum t-Test für verbundene Stichproben mit **Sig.** überschrieben. Leider gibt SPSS beim t-Test für verbundene Stichproben ausschließlich die *zweiseitige* Überschreitungswahrscheinlichkeit aus, während wir unsere allgemeinpsychologische KFA-Hypothese mit

gutem Grund einseitig formuliert haben und daher auch die einseitige Überschreitungswahrscheinlichkeit benötigen. Es wird sich gleich zeigen, dass man die einseitige Überschreitungswahrscheinlichkeit leicht aus der zweiseitigen berechnen kann.

Bei einem akzeptierten **Fehlerrisiko erster Art** von α = 5% verwendet man die folgende **Entscheidungsregel**:

$$P_{H_0}(T_Z \geq T_{emp}) \begin{cases} \geq 0{,}05 & \Rightarrow \quad H_0 \text{ beibehalten} \\ < 0{,}05 & \Rightarrow \quad H_0 \text{ verwerfen} \end{cases} \qquad (1\text{-}1)$$

Die Nullhypothese wird also abgelehnt, wenn die Teststatistik in der beobachteten Stichprobe einen Wert annimmt, der bei Zufallsstichproben aus einer H_0-Population, nur relativ selten (mit einer Wahrscheinlichkeit kleiner 0,05) erreicht oder übertroffen wird.

In Statistiklehrbüchern wird oft für den Signifikanztest zum Niveau α = 0,05 ein **kritischer Wert** T_{krit} so bestimmt, dass gilt:

$$P_{H_0}(T_Z \geq T_{krit}) = 0{,}05$$

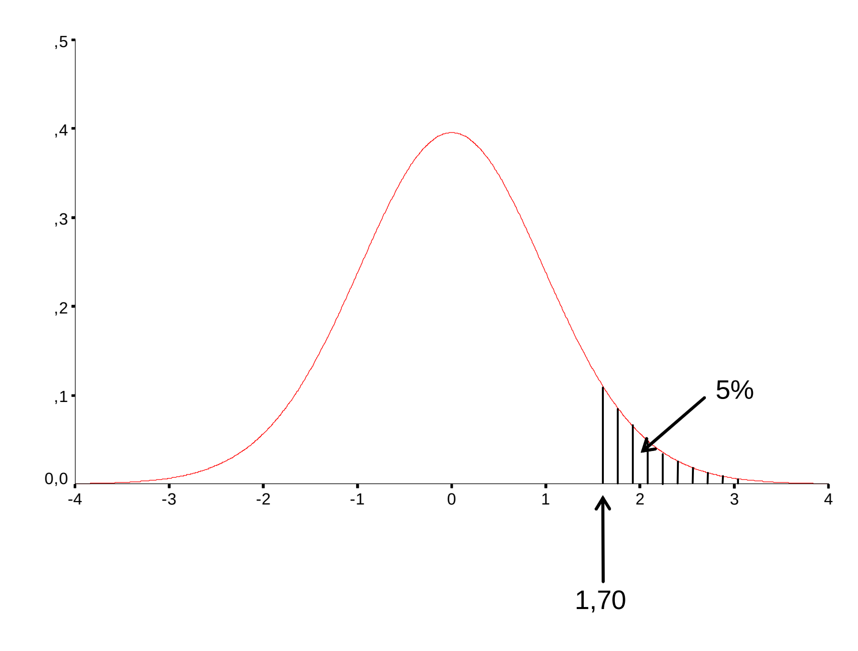
Damit kann obige Entscheidungsregel äquivalent folgendermaßen formuliert werden:

$$T_{emp} \begin{cases} \leq T_{krit} \Rightarrow & H_0 \text{ beibehalten} \\ > T_{krit} \Rightarrow & H_0 \text{ verwerfen} \end{cases} \qquad (1\text{-}2)$$

T_{krit} ist in unserer Situation gerade das 95%-Quantil der t-Verteilung mit n - 1 Freiheitsgraden. Bei der Stichprobengröße n = 31 erhalten wir T_{krit} = 1,70.

Wir haben bei den approximativen Tests in Abschnitt 4.5 die Testentscheidung anhand von kritischen Werten vorgenommen. Dort waren wir ausnahmsweise in der Lage, keine Überschreitungswahrscheinlichkeiten zu kennen, aber (Näherungen für) die kritischen Werte der Teststatistiken (als Quantile der Standardnormalverteilung) leicht ermitteln zu können. Weil SPSS und vergleichbare Statistikprogramme in der Regel Überschreitungswahrscheinlichkeiten ausgeben, werden die im Anhang vieler Statistiklehrbücher tabellierten kritischen Werte der wichtigsten Prüfverteilungen (z.B. t, F, χ^2) nur noch selten benötigt.

Die folgende Abbildung zeigt die Wahrscheinlichkeitsdichte der t-Verteilung mit 30 Freiheitsgraden und den **H_0-Ablehnungsbereich** (|||) bei einseitiger Fragestellung im Sinne unserer KFA-Hypothese:



Diese Dichte beschreibt das Verteilungsverhalten einer Zufallsgröße, zu der eine einzelne Realisation folgendermaßen zu ermitteln ist: Ziehe aus einer Population mit

$$\text{AERGZ} \sim N(0, \sigma_Z^2)$$

eine Zufallsstichprobe der Größe $n = 31$, ermittle die AERGZ-Werte und berechne T_Z. Wir kommen zu einer Testentscheidung, indem wir unser Stichprobenergebnis T_{emp} vor dem Hintergrund dieses Erwartungshorizonts beurteilen. Wir lehnen die Nullhypothese ab, wenn sie als Generator unserer Daten unplausibel ist.

Wenn wir aus einer Nullhypothesenpopulation (genauer: bei $\mu_Z = 0$) eine Zufallsstichprobe der Größe $n = 31$ ziehen und T_{emp} ermitteln, werden wir mit der Wahrscheinlichkeit $\alpha = 0{,}05$ einen Wert größer oder gleich $T_{\text{krit}} = 1{,}70$ erhalten und falsch gegen die H_0 entscheiden, also einen **Fehler erster Art** begehen. Der α-Wert sollte umso niedriger angesetzt werden, je gravierender (schädlicher, teurer) das irrtümliche Ablehnen einer gültigen Nullhypothese ist.

Das Risiko, bei Gültigkeit der *Alternativ*hypothese falsch zu entscheiden (**Fehler zweiter Art**, β-Fehler), hängt von folgenden Faktoren ab:

- **Effektstärke**

  Bei unserem KFA-Testproblem ist die Effektstärke definiert durch $dz := \frac{\mu_Z}{\sigma_Z}$ (vgl. Abschnitt 1.3.2). Je größer die Effektstärke, desto wahrscheinlicher ist ein T_{emp}-Wert jenseits von T_{krit}, also eine korrekte Entscheidung gegen die H_0.
- **Akzeptierter α-Fehler**

  Reduziert man den akzeptierten α-Fehler, so steigt der kritische Wert T_{krit}. Folglich steigt auch das Risiko dafür, dass der T_{emp} - Wert einer Stichprobe trotz gültiger Alternativhypothese den kritischen Wert T_{krit} *nicht* übertrifft. In diesem Fall kommt es *nicht* zur korrekten Entscheidung gegen die H_0, sondern zu einem Fehler zweiter Art.
- **Ein- bzw. Zweiseitigkeit des Testproblems**

  Wer sich auf eine Richtung (das Vorzeichen des Effekts) festlegt (einseitig testet), wird mit einer höheren Power belohnt (siehe unten).
- **Sensibilität des verwendeten Signifikanztests**

  Die Wahrscheinlichkeit dafür, dass ein bestimmter Populationseffekt in einer Stichprobe zu einem signifikanten Testergebnis führt, wächst mit der **Stichprobengröße**, hängt aber auch von der Güte des Verfahrens ab. Alternative Verfahren unterscheiden sich meist bei ihren Annahmen über die Skalenqualität und die Verteilung der beteiligten Variablen. In der Regel besitzt das Verfahren mit den stärksten Annahmen die beste Güte, *falls seine Voraussetzungen erfüllt sind*. Wir werden zur Prüfung der allgemeinpsychologischen Hypothese den t-Test für verbundene Stichproben nur dann einsetzen, wenn sich die Variable AERGZ in unserer Stichprobe als annähernd normalverteilt erweist. Sind die Voraussetzungen eines Verfahrens erheblich verletzt, darf es wegen potentiell verfälschter Ergebnisse nicht verwendet werden. In der Regel wäre das Verfahren in dieser Situation wegen *geringer* Sensibilität aber auch eine schlechte Wahl. Ob bereits eine erhebliche Verletzung der Voraussetzungen vorliegt, oder noch auf die Robustheit eines Verfahrens vertraut werden kann, ist leider oft schwer zu entscheiden.

Wie Sie aus der Stichprobenumfangsplanung in Abschnitt 1.3.2 wissen, kann man zum t-Test für abhängige Stichproben für eine konkret vorgegebene Effektstärke *dz*, eine Testausrichtung (ein- oder zweiseitig) und ein α-Fehlerniveau …

- die Teststärke (Power) bzw. das β-Fehler-Risiko zu einer festen Stichprobengröße ausrechnen,
- für eine erwünschte Teststärke (z.B. $1 - \beta = 0{,}8$) die erforderliche Stichprobengröße ermitteln.

Passend zu unserer allgemeinpsychologischen KFA-Hypothese haben wir bislang das *einseitige* Testproblem behandelt. Wir wollen noch das folgende **zweiseitige Testproblem** betrachten:

$$\mathrm{H}_0 : \mu_M = \mu_O \quad \text{vs.} \quad \mathrm{H}_1 : \mu_M \neq \mu_O$$

bzw.

$$\mathrm{H}_0 : \mu_Z = 0 \quad \text{vs.} \quad \mathrm{H}_1 : \mu_Z \neq 0$$

Die H_0 des zweiseitigen Tests ist gerade identisch mit dem *Rand* der H_0 zum einseitigen Test.

Wir verwenden beim zweiseitigen Test dieselbe Teststatistik T_Z wie beim einseitigen Test. Nun sind aber *betragsmäßig* große T_{emp}-Werte (mit *positivem oder negativem* Vorzeichen) indikativ für eine Abweichung von der Nullhypothese. Nach einem generellen Prinzip der Testkonstruktion müssen *alle* Elemente der Alternativhypothese (im zweiseitigen Fall also mit $\mu_Z < 0$ oder $\mu_Z > 0$) eine faire Chance haben, sich in einem signifikanten Ergebnis zu artikulieren. Anderenfalls resultiert ein so genannter *verfälschter Test*. Daher ist die *zweiseitige* Überschreitungswahrscheinlichkeit

$$\mathrm{P}_{\mathrm{H}_0}(|T_Z| \geq |T_{\mathrm{emp}}|)$$

zu ermitteln und in folgender Entscheidungsregel zu verwenden:

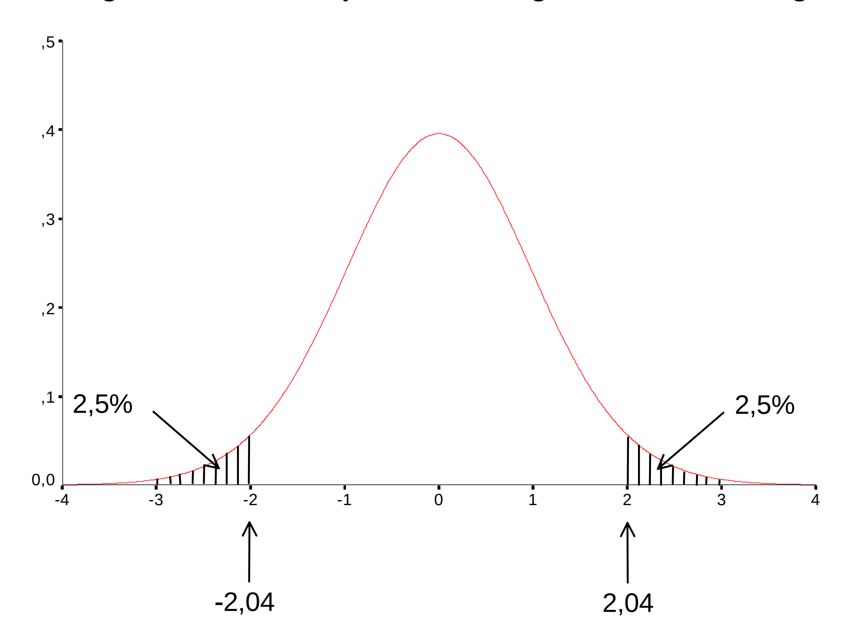
$$\mathrm{P}_{\mathrm{H}_0}(|T_Z| \geq |T_{\mathrm{emp}}|) \begin{cases} \geq 0{,}05 & \Rightarrow \quad \mathrm{H}_0 \text{ beibehalten} \\ < 0{,}05 & \Rightarrow \quad \mathrm{H}_0 \text{ verwerfen} \end{cases} \tag{1-3}$$

Der kritische Werte $T_{\mathrm{krit},2}$ zum zweiseitigen Test zum Niveau $\alpha = 0{,}05$ ist so zu bestimmen, dass gilt:

$$\mathrm{P}_{\mathrm{H}_0}(|T_Z| \geq |T_{\mathrm{krit},2}|) = 0{,}05$$

Bei unserer Stichprobengröße $n = 31$ erhalten wir $T_{\mathrm{krit},2} = \pm 2{,}04$.

Bei zweiseitiger Testung haben wir zwei symmetrisch angeordnete Ablehnungsbereiche:



Weil unsere Teststatistik symmetrisch um den Wert Null verteilt ist, gilt für $T_{emp} \geq 0$:

$$P_{H_0}(T_Z \geq T_{emp}) = \frac{1}{2} \cdot P_{H_0}(|T_Z| \geq |T_{emp}|) \quad (1\text{-}4)$$

Die Überschreitungswahrscheinlichkeit des einseitigen t-Tests ergibt sich also durch Halbieren aus der Überschreitungswahrscheinlichkeit des zweiseitigen t-Tests (, sofern die Prüfgröße das von der H_1 behauptete Vorzeichen besitzt). Dieser Zusammenhang ist wichtig in der statistischen Praxis mit SPSS, weil dieses Programm bei t-Tests häufig nur die zweiseitige Überschreitungswahrscheinlichkeit mitteilt. Sie dürfen aber den Zusammenhang in Gleichung (1-4) keinesfalls auf beliebige Tests generalisieren. Wir werden z.B. im Zusammenhang mit der Kreuztabellenanalyse (siehe Abschnitt 11.4.4.1) den exakten Test von Fisher kennen lernen, bei dem eine analoge Gleichung *nicht* gilt.

7.2 Zu den Voraussetzungen der zentralen Hypothesentests

Der t-Test für verbundene Stichproben, mit dem wir unsere *allgemeinpsychologische* Hypothese prüfen wollen, setzt voraus, dass die Differenzvariable AERGZ normalverteilt ist (vgl. Abschnitt 7.1). Diese Normalverteilungsannahme soll anschließend mit der SPSS-Prozedur zur explorativen Datenanalyse geprüft werden.

Unsere *differentialpsychologische* Hypothese bezieht sich auf den Steigungskoeffizienten β_1 in der linearen Regression von AERGAM auf LOT:

$$\text{AERGAM} = \beta_0 + \beta_1\text{LOT} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Die Hypothesen des Testproblems lauten:

$$H_0: \beta_1 \geq 0 \quad \text{vs.} \quad H_1: \beta_1 < 0$$

Es kommt eine Teststatistik zum Einsatz, die sich im vorliegenden Fall der *bivariaten* Regression besonders bequem mit Hilfe der Stichprobenkorrelation *r* zwischen Kriterium und Regressor notieren lässt:

$$T_r := \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Als Effektstärkemaß zur differentialpsychologischen Hypothese haben wir in Abschnitt 1.3.3 kennen gelernt:

$$f^2 = \frac{R^2}{1-R^2}$$

Offenbar ist das Quadrat der Prüfgröße

$$T_r^2 = \frac{r^2}{1-r^2}(n-2)$$

abgesehen vom (datenunabhängigen) Faktor (n - 2) ein Schätzer für die Teststärke. Damit zeigt die Prüfgröße T_r (bei passendem Vorzeichen) an, wie stark die Daten von der Nullhypothesen-Behauptung abweichen. Außerdem ist T_r bei gültiger Nullhypothese (genauer: bei $\beta_1 = 0$) t-verteilt mit n - 2 Freiheitsgraden, sofern die Voraussetzungen des Regressionsmodells erfüllt sind, die anschließend der bequemeren Schreibweise halber für ein Kriterium *Y* und einen Regressor *X* beschrieben werden:

1) **Linearität**

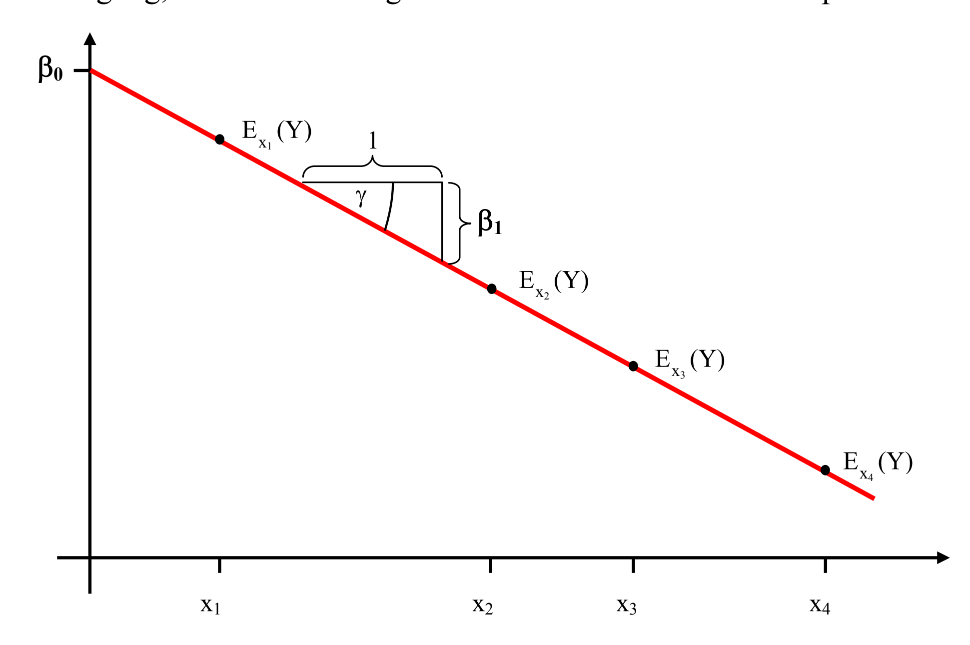
Der Erwartungswert (Mittelwert) $E_X(Y)$ von *Y* für einen bestimmten *X*-Wert hängt **linear** von *X* ab:

$$E_X(Y) = \beta_0 + \beta_1 X$$

Für beliebige *X*-Ausprägungen liegen die zugehörigen Erwartungswerte $E_X(Y)$ auf der **Regressionsgeraden** durch die Punktepaare

$$(X, \beta_0 + \beta_1 X)$$

Dabei ist β_0 der Schnittpunkt der Regressionsgeraden mit der Y-Achse (Ordinatenabschnitt) und β_1 das Gefälle der Regressionsgeraden (der Tangens des Winkels γ der Regressionsgeraden mit der X-Achse). Unserer differentialpsychologischen Hypothese entspricht eine Regressionsgerade mit negativer Steigung, weil wir eine Ärgerreduktion bei zunehmendem Optimismus erwarten.



Zur Interpretation des Koeffizienten β_1: Erhöht man *X* um eine Einheit, so sinkt gemäß modellgemäß der Erwartungswert $E_X(Y)$ um β_1 Einheiten (negative Steigung im Sinn der Alternativhypothese).

2) **Normalität der Residuen**

Für die (nicht direkt beobachtbare) Fehler- bzw. Residualvariable ε wird angenommen, dass sie **normalverteilt** ist mit Erwartungswert Null und Varianz σ^2. Sie dürfen sich vorstellen, dass es für jede *X*-Ausprägung eine **Normalverteilung** potentieller ε-Werte gibt, aus der **zufällige** Realisationen gezogen werden, die zusammen mit dem konstanten Anteil $\beta_0 + \beta_1 X$ die Realisationen der abhängigen Variablen *Y* ergeben.

3) **Varianzhomogenität der Residuen (Homoskedastizität)**

Die Normalverteilungen der ε-Variablen zu den verschiedenen *X*-Ausprägungen haben alle **dieselbe Varianz** σ^2.

4) **Unabhängigkeit der Residuen**

Die Residuen zu den einzelnen Beobachtungen (Fällen) in der Stichprobe sind unkorreliert. Wegen ihrer Normalverteilung sind sie damit auch stochastisch unabhängig.

Hinsichtlich der Verteilungsvoraussetzungen ist zu betonen:

- Es wird keine Annahme über die Verteilung des Regressors gemacht.
- Es wird keine Annahme über die *univariate* Verteilung des Kriteriums (die so genannte Randverteilung) gemacht.
- Es sind die **Residuen des Modells**, die bestimmte Verteilungsvoraussetzungen erfüllen müssen (Erwartungswert Null, Normalität, Homoskedastizität, Unabhängigkeit).

Für methodisch besonders Interessierte soll noch eine alternative Darstellung für T_r vorgeführt werden, die von eher anwendungsorientierten Lesern gefahrlos übersprungen werden kann. Weil der Stichprobenschätzer b_1 des Steigungskoeffizienten in folgender Beziehung zur Stichprobenkorrelation r und den Schätzern S_Y und S_X für die Standardabweichungen des Kriteriums Y und des Regressors X steht

$$b_1 = r\,\frac{S_Y}{S_X}$$

und der geschätzte Standardfehler zu b_1 gleich

$$sf_{b_1} = \frac{S_Y}{S_X}\frac{\sqrt{1-r^2}}{\sqrt{n-2}}$$

ist (siehe z.B. Cohen et al. 2003, S. 42), kann auch die Prüfgröße T_r als Quotient aus einem Stichprobenschätzer und seinem geschätzten Standardfehler geschrieben werden:

$$T_r = b_1\,\frac{S_X}{S_Y}\frac{\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{b_1}{sf_{b_1}}$$

7.3 Verteilungsanalyse für die abgeleiteten Variablen

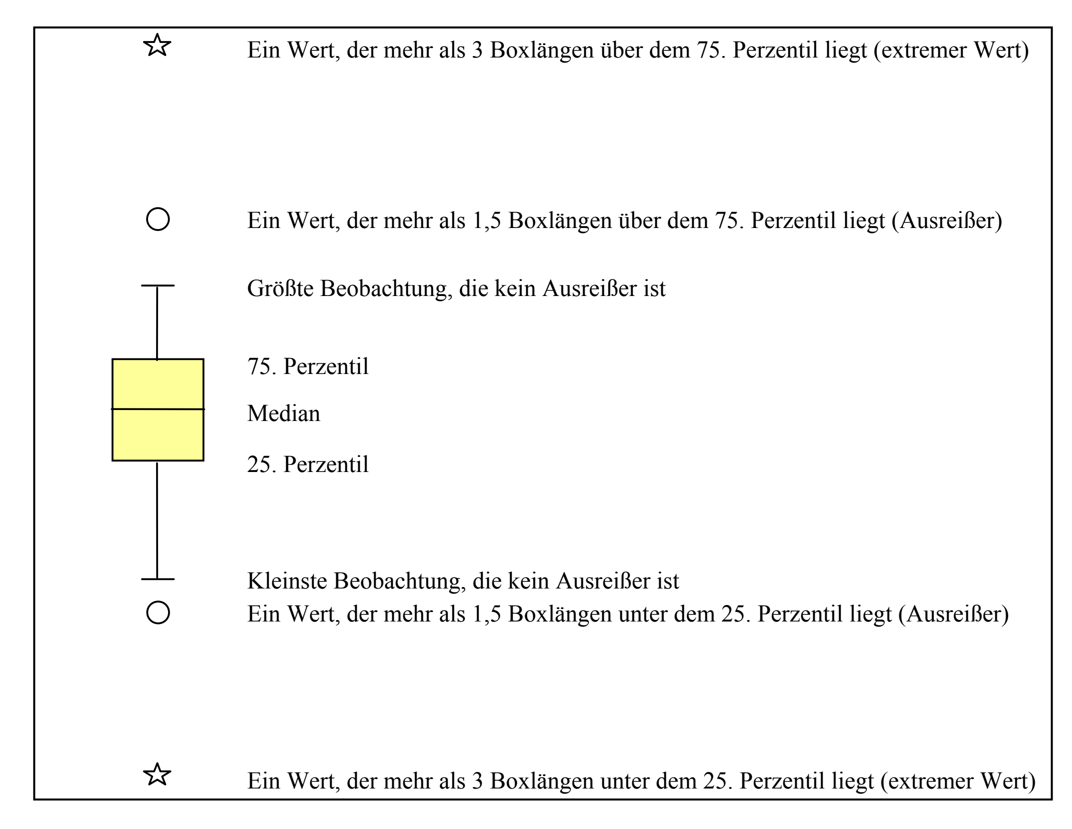
Für die folgenden Schritte wird eine aktive SPSS-Sitzung mit geöffneter Fertigdatendatei **kfa.sav** vorausgesetzt. Ob Sie die SPSS-Kommandos zu den anstehenden Analysen für spätere Wiederverwendung konservieren wollen, bleibt Ihnen überlassen.

Wir wollen zunächst die univariaten Verteilungen der abgeleiteten Variablen LOT, AERGAM, AERGZ und BMI untersuchen. Analog zu den Verteilungsanalysen in Abschnitt 4, die auch zur Datenprüfung dienten, wollen wir bei den Verteilungen der abgeleiteten Variablen auch auf Anomalien infolge fehlerhafter oder schlecht durchdachter Berechnungsvorschriften achten. Außerdem wollen wir noch eine weitere Gefahrenquelle für unser Forschungsprojekt ins Visier nehmen:

7.3.1 Diagnose von Ausreißern

Als **Ausreißer** bezeichnet man extreme Werte, die zwar innerhalb des logisch möglichen Wertebereichs liegen, aber doch mit großer Wahrscheinlichkeit nicht aus der interessierenden Verteilung bzw. Population stammen. Diese Werte haben insbesondere auf parametrische Auswertungsverfahren einen starken, verzerrenden Einfluss. Daher wollen wir ab jetzt auch auf Ausreißer achten.

Dazu lassen wir uns für jede Variable einen **Boxplot** erstellen. Dieses beliebte Instrument der explorativen Datenanalyse zeigt auf prägnante Weise wesentliche Verteilungsinformationen, und ist zur Identifikation von Ausreißern gut geeignet. Die Bestandteile eines Boxplots haben folgende Bedeutung:



Als Ursachen für Ausreißer kommen in Frage:

- Erhebungs- bzw. Erfassungsfehler
  Messwerte können falsch ermittelt oder fehlerhaft in die EDV übernommen worden sein.
- Besondere Umstände beim Merkmalsträger
  Bei einer Agrarstudie zum Ertrag verschiedene Getreidesorten kann z.B. der Boden in einer bestimmten Versuchsparzelle durch einen Ölunfall verseucht worden sein.

Eindeutig irreguläre Daten müssen natürlich entfernt werden. Sie können z.B. mit dem Dateneditor in der Rohdatendatei:

- einen Wert löschen, d.h. durch SYSMIS ersetzen
- einen Wert als MD-Indikator deklarieren
- einen kompletten Fall löschen

Natürlich dürfen Sie keine Daten eliminieren, weil sie Ihren Hypothesen widersprechen.

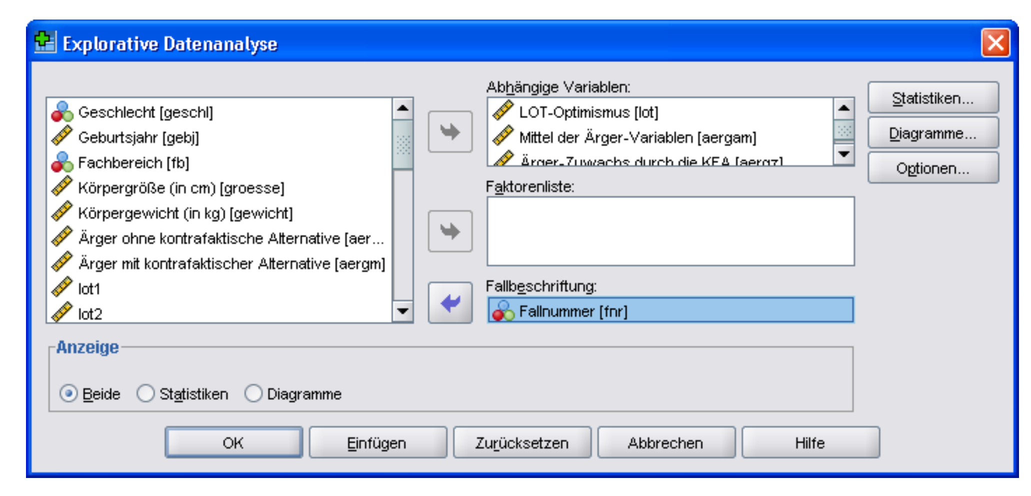
7.3.2 Die SPSS-Prozedur zur explorativen Datenanalyse

Für die eben geplanten Aufgaben (Ausreißerdiagnose und Verteilungsprüfung) eignet sich die SPSS-Prozedur zur explorativen Datenanalyse besser als die in Abschnitt 4 der Einfachheit halber bevorzugte Häufigkeitsanalyse. Natürlich können Sie in Zukunft auch die Verteilungen von Rohvariablen mit der leistungsfähigeren explorativen Datenanalyse untersuchen.

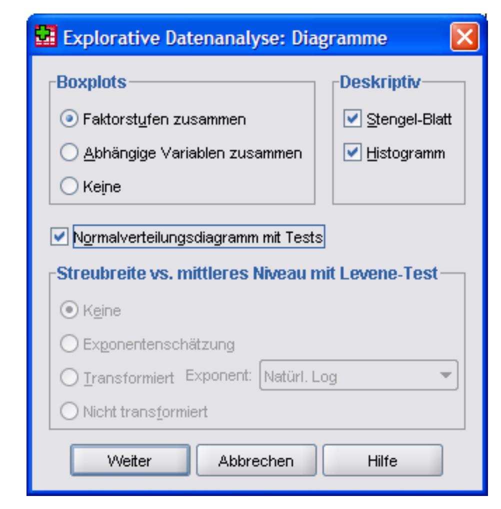
Starten Sie deren Dialogbox mit:

Analysieren > Deskriptive Statistiken > Explorative Datenanalyse

Transportieren Sie die Namen der drei zu untersuchenden Variablen in die Liste der **abhängigen Variablen**, und wählen Sie die Variable FNR zur Fallbeschriftung aus, damit mögliche Ausreißer durch ihre Fallnummer identifiziert werden können:



Fordern Sie in der **Diagramme**-Subdialogbox zusätzlich **Histogramme** sowie **Normalverteilungdiagramme mit Tests** an:



Das Kontrollkästchen zum Anfordern von Normalverteilungsanpassungstests (Kolmogorov-Smirnov und Shapiro-Wilk) hat SPSS wirklich sehr gut in der **Diagramme**-Subdialogbox der explorativen Datenanalyse versteckt.

Der Klarheit halber soll nochmals betont werden, dass wir nur für die Variable AERGZ eine Normalverteilungsvoraussetzung zu prüfen haben (vgl. Abschnitt 7.2). Allerdings sind die teilweise irrelevanten Ausgaben für LOT, AERGAM und BMI kein Grund dafür, zwei verschiedene Analysen anzufordern.

Wir erhalten im Ausgabefenster u.a. für jede abhängige Variable einen **Boxplot**.

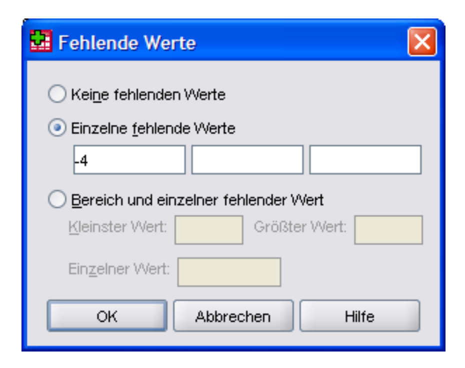
7.3.3 Ergebnisse für AERGZ

Bei der Ausreißeranalyse gibt es nur einen Problemfall und zwar ausgerechnet bei der Variablen AERGZ, über die unsere zentrale KFA-Hypothese geprüft werden soll. Hier tanzt Fall Nr. 4 aus der Reihe:



Diese Person hatte *ohne* KFA eine Ärgertemperatur von 60° gemeldet, die sich dann durch die KFA-Komponente auf 20° abkühlte. Zwar darf dieses Muster nicht a-priori als verdächtig gelten, weil es unserer Hypothese widerspricht, doch der Boxplot gibt eine klare Empfehlung, den Fall bei dieser Analyse auszuschließen. Allerdings scheut sich ein redlicher Forscher, Daten zu neutralisieren, die der eigenen Hypothese widersprechen.

Vor einer endgültigen Entscheidung wollen wir die Verteilung von AERGZ noch weiter analysieren, da beim geplanten t-Test zur allgemeinpsychologischen KFA-Hypothese vorausgesetzt werden muss, dass AERGZ (in der Population) normalverteilt ist. Damit der extreme AERGZ-Wert von Fall Nr. 4 die weitere Verteilungsanalyse nicht beeinflusst, soll er vorübergehend neutralisiert werden. Weil wir noch keine Methode kennen, komplette Fälle von einer Analyse fern zu halten (siehe Abschnitt 10), deklarieren wir den betroffenen Wert (= -4) als MD-Indikator. Auf diese Weise findet sich doch noch eine Gelegenheit, die Deklaration von benutzerdefinierten MD-Indikatoren zu üben. Markieren Sie in der Variablenansicht des Datenfensters die Zelle mit den **Fehlenden Werten** der Variablen AERGZ, klicken Sie auf den Erweiterungsschalter ..., und tragen Sie den Wert –4 als einzelnen MD-Indikator ein:



Das folgende Histogramm zeigt, dass die AERGZ-Verteilung *auch nach* Elimination von Fall Nr. 4 noch relativ deutlich von der Normalität abweicht:



Tatsächlich lehnen auch *nach* der Elimination des Ausreißers die beiden von SPSS angebotenen Normalverteilungstests (Kolmogorov-Smirnov und Shapiro-Wilk) die im t-Test benötigte Normalverteilungsannahme ab:

Tests auf Normalverteilung

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
Ärger-Zuwachs durch die KFA	,207	30	,002	,913	30	,018

a. Signifikanzkorrektur nach Lilliefors

Auch diese Testentscheidung folgt der in Abschnitt 7.1 beschriebenen Logik, wobei folgende Hypothesen zur Konkurrenz stehen:

H_0: AERGZ ist normalverteilt versus H_1: AERGZ ist *nicht* normalverteilt

Die von SPSS berechnete Überschreitungswahrscheinlichkeit (**Signifikanz**) ist bei beiden Teststatistiken kleiner als 0,05, so dass beide Tests übereinstimmend die Nullhypothese verwerfen. Dies ist vor allem deshalb ein ernst zu nehmender Befund, weil unsere Stichprobe relativ klein und damit die Power der Tests eher gering ist.

Bei einer *großen* Stichprobe besitzen die Normalitätstests eine hohe Power und decken auch kleine (für die Validität des geplanten t-Tests irrelevante) Abweichungen von der Nullhypothese auf. Folglich ist dann ein signifikantes Testergebnis „nicht tragisch“. Wenn bei einer *kleinen* Stichprobe ein Normalitätstest „anschlägt“, ist jedoch von einer relevanten Verletzung der Normalitätsannahme auszugehen.

Aufgrund der problematischen Verteilungsverhältnisse entscheiden wir uns, statt des geplanten parametrischen t-Tests für verbundene Stichproben einen verteilungsfreien Lagevergleich mit dem **Vorzeichentest** durchzuführen (siehe z.B. Hartung 1989, S. 242f). Dieser Test entscheidet sich zwischen folgenden Hypothesen:

H_0: Der Median der Differenzvariablen AERGZ ist kleiner oder gleich Null.

versus

H_1: Die Differenzvariable AERGZ hat einen positiven Median. (Mehr als 50% der Fälle haben einen positiven AERGZ-Wert.)

Statt der in Abschnitt 7.1 ausführlich vorgestellten Teststatistik T_Z verwendet der Vorzeichentest eine Prüfgröße, die im Wesentlichen auf der Anzahl der positiven AERGZ-Ausprägungen in der Stichprobe basiert. Sie wird üblicherweise mit Z bezeichnet, weil sie unter der H_0 (genauer: bei einem Median von Null) approximativ z-verteilt (d.h. standardnormalverteilt) ist. Leider kollidiert die Bezeichnung mit der oben eingeführten Abkürzung für unsere Ärgerzuwachsvariable.

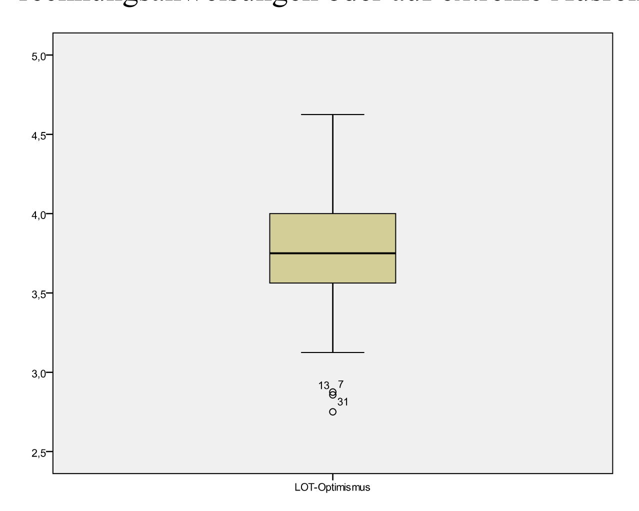
Man geht davon aus, dass die Verteilungs-Approximation ab $n \geq 20$ hinreichend genau ist, so dass wir den Test bei unserer Stichprobe ($n = 31$) in der üblichen approximativen Form anwenden dürfen. Bei kleineren Stichproben muss die *exakte* Variante des Tests eingesetzt werden, die von SPSS ebenfalls unterstützt wird.

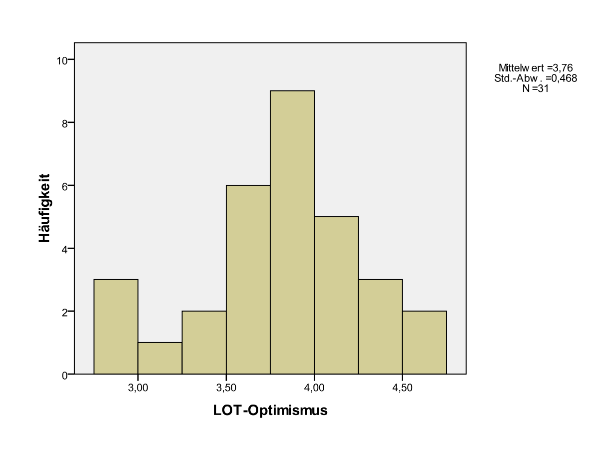
Weil der Vorzeichentest weit weniger empfindlich auf Ausreißer reagiert als der parametrische t-Test, können wir den kritischen Fall Nr. 4 in der Auswertung belassen. Damit vermeiden wir den Verdacht, die Daten zu unseren Gunsten bereinigt zu haben. Heben Sie also bitte die MD-Deklaration für den Wert –4 bei der Variablen AERGZ wieder auf.

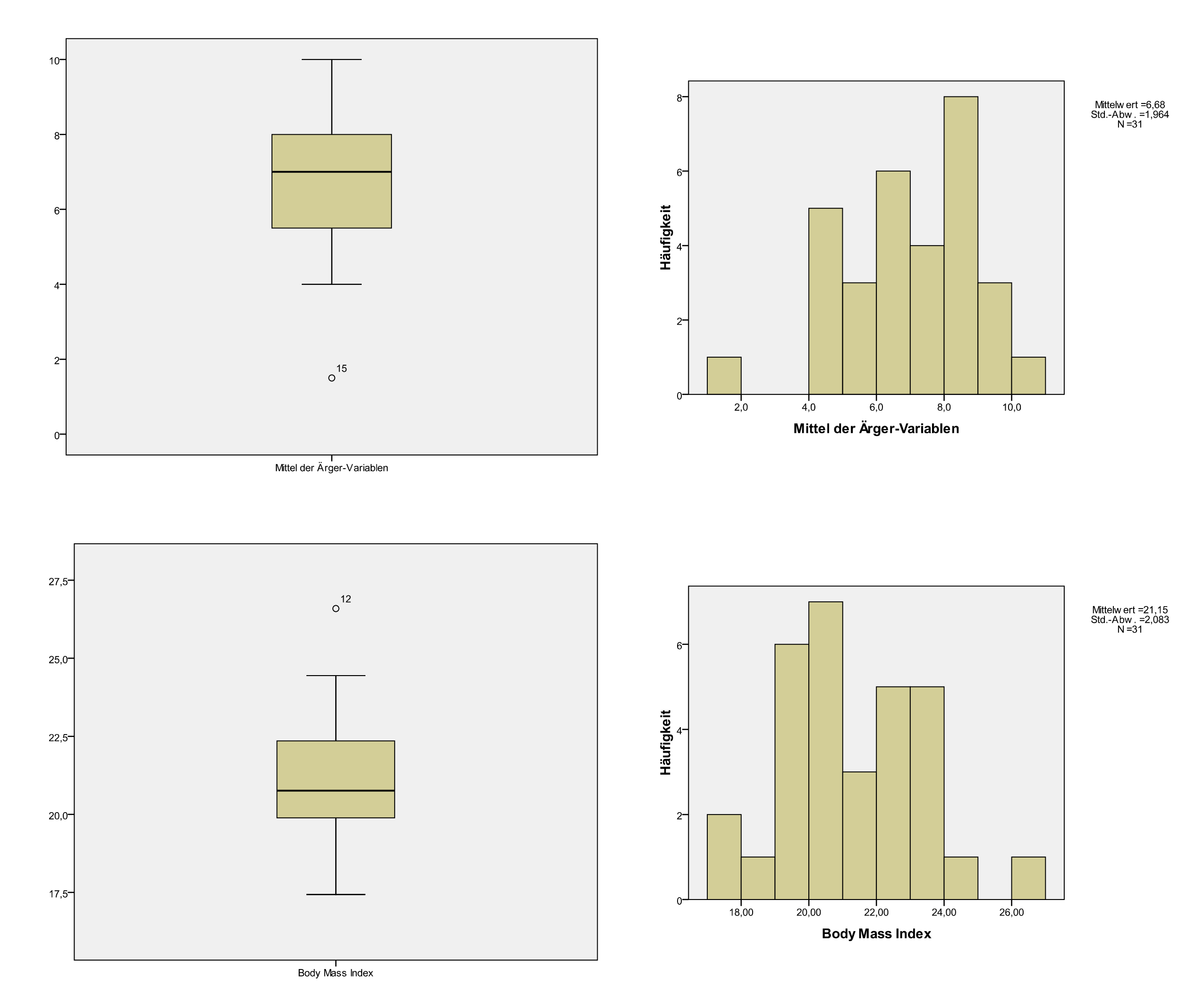
Die bisherige Diskussion der AERGZ-Verteilung hat sich auf Gefahrenquellen für die Interpretierbarkeit des geplanten zentralen Hypothesentests konzentriert. Es ist jedoch keinesfalls verboten, sondern sogar dringend empfohlen, sich anhand obiger Verteilungsdiagramme und sonstiger deskriptiver Informationen einen Eindruck von der empirischen Bewährung der KFA-Hypothese zu verschaffen. Das AERGZ-Histogramm spricht für einen starken KFA-Effekt in der erwarteten Richtung. Eine genaue Kenntnis des deskriptiven Ergebnisbilds kann verhindern, dass wir von einem durch technische Defekte verfälschten Testergebnis in die Irre geführt werden.

7.3.4 Ergebnisse für LOT, AERGAM und BMI

Bei den Variablen LOT, AERGAM und BMI finden sich keine Hinweise auf Fehler in den Berechnungsanweisungen oder auf extreme Ausreißer:







Die in den Boxplots auftauchenden Ausreißer sind nicht extrem (Abstand vom 25. Perzentil kleiner als drei Boxlängen), und sollten aufgrund einer relativ kleinen Stichprobe, welche die Populationsverteilungen nur grob charakterisiert, *nicht* ausgeschlossen werden.

Bei der geplanten Regressionsanalyse mit AERGAM und LOT hat zudem die Ausreißeranalyse auf der Basis der Modellresiduen das weit größere Gewicht.

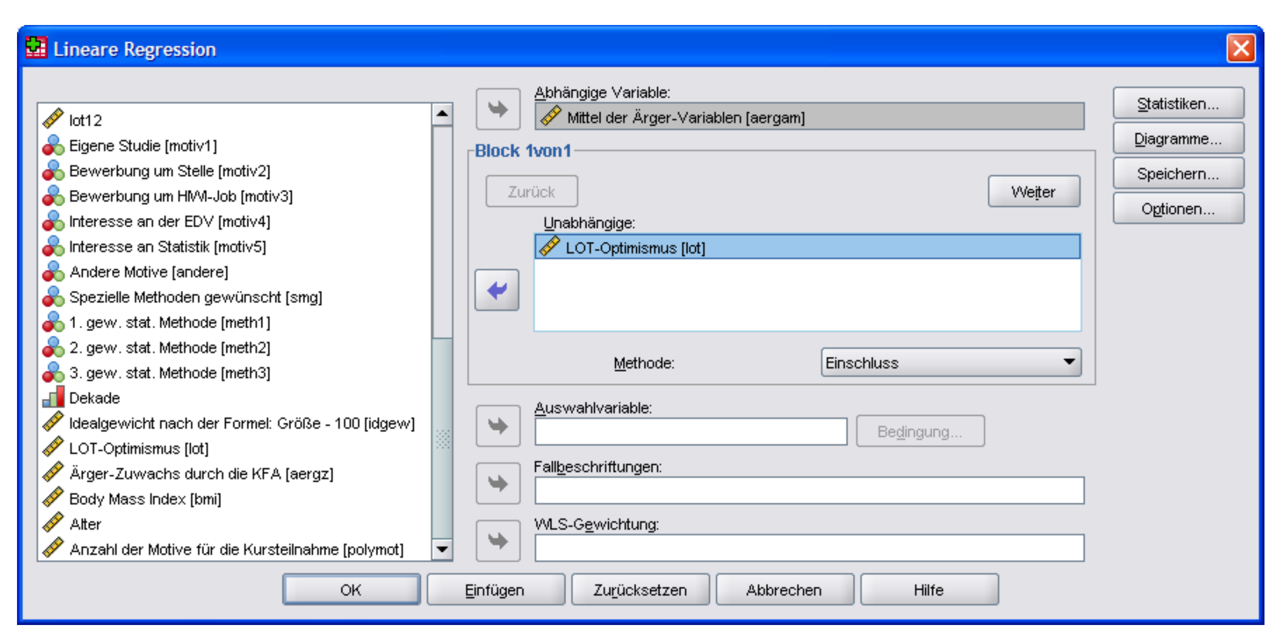
7.4 Prüfung der differentialpsychologischen Hypothese

7.4.1 Regression von AERGAM auf LOT

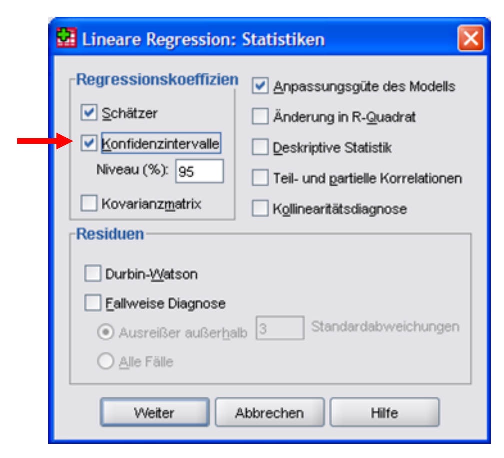
Nun wollen wir die lineare Regression von AERGAM auf LOT untersuchen, die wir nach dem Menübefehl

Analysieren > Regression > Linear

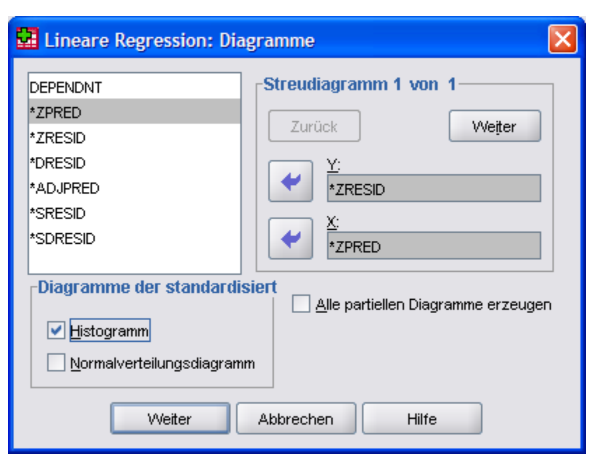
in der folgenden Dialogbox anfordern können:



In der **Statistiken** - Subdialogbox verlangen wir über die Voreinstellung hinausgehend die Berechnung von Vertrauens- bzw. Konfidenzintervallen:



Zur Prüfung der in Abschnitt 7.2 beschriebenen Voraussetzungen ordern wir in der **Diagramme**-Subdialogbox

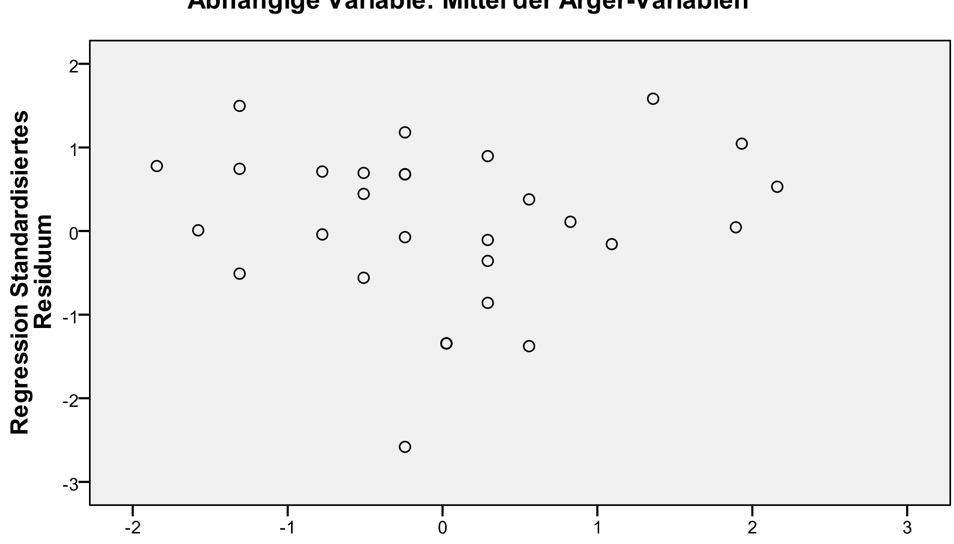


folgende Ausgaben:

- Das **Streudiagramm** der standardisierten Residuen gegen die standardisierte Modellprognose
  Für jeden prognostizierten Wert (also letztlich für jeden Wert des Regressors) sollten sich die Residuen varianzhomogen um den Erwartungswert Null verteilen.

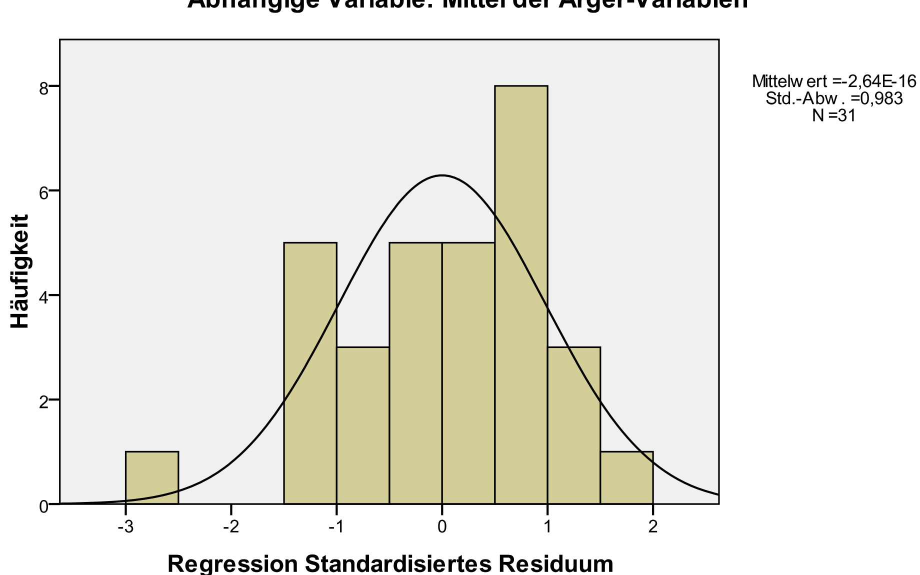
- Das **Histogramm** der standardisierten Residuen
  Der geplante Signifikanztest zum Steigungskoeffizienten der linearen Regression setzt normalverteilte Residuen voraus.

Das Streudiagramm bietet wenig Anlass zur Sorge um die Linearität und die Homoskedastizität:

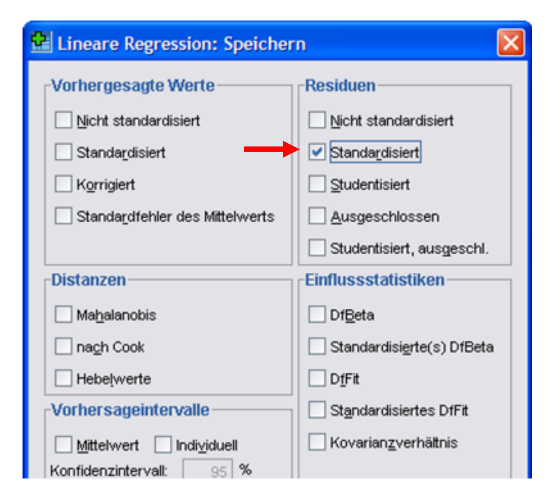


Wir sehen ein „signifikantes" Residuum (standardisierter Wert betragsmäßig größer Zwei), was aber bei 31 Fällen mit der Annahme eines gültigen Modells vereinbar ist.

Das Histogramm der standardisierten zeigt sich eine zufriedenstellende Normalverteilungsapproximation:



Mit den per **Speichern**-Subdialog



in eine neue Variable geschriebenen Residuen lässt sich auch ein formaler Normalverteilungsanpassungstest durchführen (vgl. Abschnitt 7.3.3), doch führen derartige Voraussetzungsprüfungen per Signifikanztest nicht unbedingt auf einfache Weise zu einer guten Entscheidung, denn:

- Bei einer kleinen Stichprobe sind Verletzungen der Normalität ernst zu nehmen, können aber mangels Teststärke schwer nachgewiesen werden.
- Bei einer großen Stichprobe verliert die Normalitätsannahme an Bedeutung (zentraler Grenzwertsatz), doch werden hier auch kleine (und für die geplante Inferenzstatistik irrelevante) Abweichungen von der idealen Glockenform signifikant.

In unserem Beispiel übersteht die Annahme normalverteilter Residuen auch die Signifikanztests nach Kolmogorov-Smirnov bzw. Shapiro-Wilk:

Tests auf Normalverteilung

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
Standardized Residual	,114	31	,200*	,950	31	,154

a. Signifikanzkorrektur nach Lilliefors

*. Dies ist eine untere Grenze der echten Signifikanz.

Nachdem wir die Voraussetzungen als gültig akzeptiert haben, steht einer Inspektion der Regressionsergebnisse nichts mehr im Wege. Wir erhalten zwar, wie erwartet, einen negativen Regressionskoeffizienten, doch ist dieser bei weitem nicht signifikant:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressionskoeffizient B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	7,669	2,947		2,602	,014	1,641	13,697
	LOT-Optimismus	-,264	,778	-,063	-,339	,737	-1,854	1,327

a. Abhängige Variable: Mittel der Ärger-Variablen

SPSS ermittelt eine zweiseitige Überschreitungswahrscheinlichkeit von 0,737, die auch nach der zulässigen Halbierung aufgrund unserer einseitigen Fragestellung von der Signifikanzgrenze 0,05 sehr weit entfernt ist. Der LOT-Optimismus zeigt entgegen unserer Annahme fast keinen linearen Zusammenhang mit dem mittleren Ärger in unserer fiktiven Situation.

Wer sich ausführlich über die lineare Regressionsanalyse mit SPSS informieren möchte, kann eine elektronische Publikation des Rechenzentrums zu diesem Thema (Baltes-Götz 2008a) auf dem Webserver der Universität Trier von der Startseite (http://www.uni-trier.de/) ausgehend folgendermaßen finden:

Rechenzentrum > Studierende > EDV-Dokumentationen >
Statistik > Lineare Regressionsanalyse mit SPSS

7.4.2 Methodologische Anmerkungen

7.4.2.1 Explorative Analysen im Anschluss an einen „gescheiterten" Hypothesentest

Auf das „Scheitern" einer konfirmatorischen Forschungsbemühung werden in der Regel explorative Analysen folgen, wobei revidierte bzw. neue Hypothesen entstehen können. Wir werden uns in Abschnitt 9.4 z.B. dafür interessieren, ob eventuell das Geschlecht den Zusammenhang zwischen Optimismus und Ärger moderiert. Allerdings ist es *nicht* möglich, revidierte oder neue Hypothesen anhand *derselben* Stichprobe zu testen. Sie dürfen und sollen aus Ihren Daten etwas lernen, aber ein Test der dabei generierten Hypothesen erfordert eine neue, unabhängige Stichprobe.

Außerdem sollten Sie es nicht unterlassen, das „Scheitern" einer Hypothese zu veröffentlichen. Ansonsten tragen Sie dazu bei, in der Fachliteratur ein systematisch verzerrtes Bild der Wirklichkeit aufzubauen.

7.4.2.2 Post hoc - Poweranalyse

Bei der Interpretation des obigen Resultats ist außerdem zu beachten, dass die Power des t-Tests zum Regressionskoeffizienten in unserer relativ kleinen Stichprobe recht bescheiden ist, so dass kleine Effekte leicht übersehen werden können. Unser Testergebnis kann nicht als *Beleg* für die Nullhypothese interpretiert werden, doch spricht es gegen die Existenz eines *starken* Effekts. Um zu genaueren Aussagen zu kommen, betrachten wir die Power unseres t-Tests bei unterschiedlichen Effektstärken in der Population.

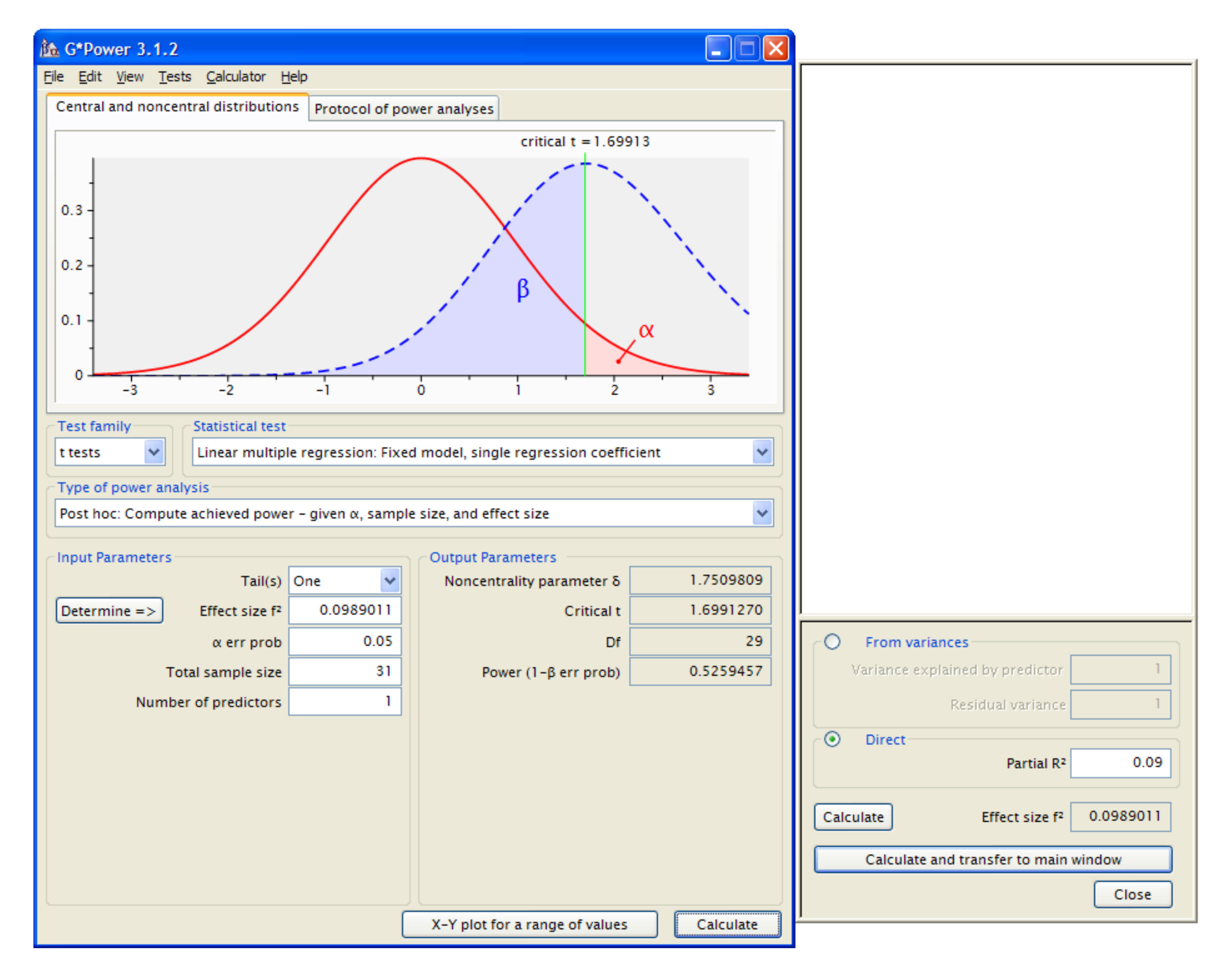
Dabei verwenden wir erneut das Programm **G*Power 3.1**, das schon bei der Stichprobenumfangsplanung in Abschnitt 1.3 zum Einsatz kam. Auf den Pool-PCs der Universität Trier unter dem Betriebssystem Windows ist G*Power 3.1 folgendermaßen zu starten

Start > Programme > Wissenschaftliche Programme > GPower

Wir wählen:

- **Test family**: **t-Tests**
- **Statistical test**: **Linear Multiple Regression: Fixed model, single regression coefficient**
- **Type of power analysis** **Post hoc**
- **Tail(s)** **One**
- **Effect size f^2** 0.0989011
- **α err prob** 0.05
- **Total sample size** 31
- **Number of predictors** 1

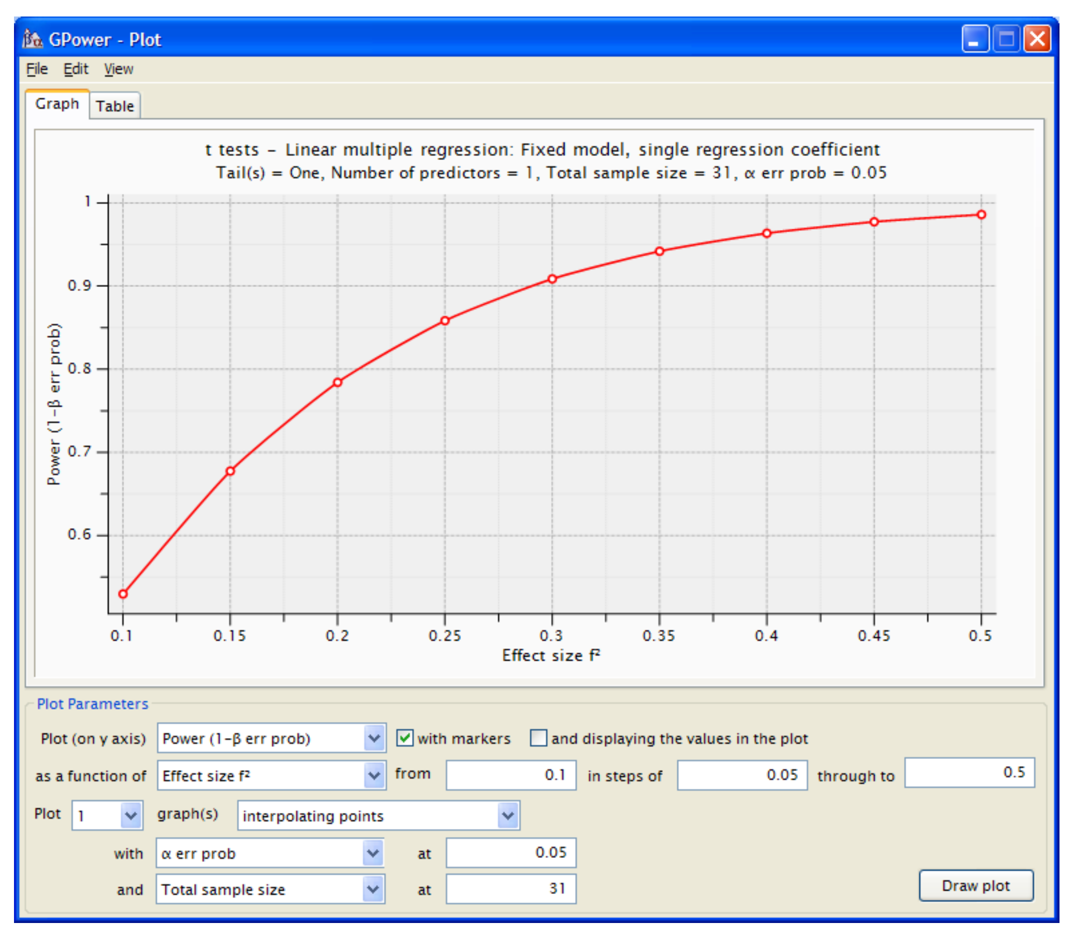
Die Effektstärke von ca. 0,1 resultiert aus der Annahme, dass 9 % der Kriteriumsvarianz durch den Regressor aufgeklärt werden können. Nach einem Klick auf den Schalter **Calculate** wird für den Test zur differentialpsychologischen Hypothese eine Teststärke (Power) von lediglich 0,53 berechnet:



Um zu einer Darstellung der Power als Funktion der Effektstärke zu gelangen, klicken wir auf den Schalter **X-Y plot for a range of values** und wählen

- **Plot (on y axis)** **Power (1 - β err prob)**
- **as a function of** **Effect size f^2**
- **from** 0.0
- **in steps of** 0.05
- **through to** 0.5
- **Plot** 1

Nach einem Klick auf den Schalter **Draw Plot** zeigt die folgende Abbildung, wie bei fester Stichprobengröße (n = 31) die Power des einseitigen Tests von der Effektstärke abhängt:



Erst ab einer Effektstärke von ca. f^2 = 0,35 (bzw. R^2 = 0,26) ist die Power so groß (ca. 0,95), dass man die ausgebliebene Signifikanz als Beleg gegen einen Effekt dieser Stärke werten kann. Unserer Studie hat also keinesfalls die differentialpsychologische Nullhypothese bewiesen, aber doch ein Argument gegen die Existenz eines starken Effekts ($f^2 \geq 0{,}35$) geliefert.

7.4.2.3 Fehlende Werte

Fehlende Werte haben Einbußen bei der Teststärke und oft auch verzerrte Schätzwerte zur Folge, so dass einige Anstrengungen zur Vermeidung oder Reduktion des Problems angemessen sind. Wir haben bei der Berechnung des LOT-Werts geeignete Maßnahmen ergriffen, um die Anzahl fehlender Werte gering zu halten (vgl. Abschnitt 6.4).

Wer sich über die in SPSS und im Strukturgleichungsanalyseprogramm Amos enthaltenen Möglichkeiten zur Analyse und Behandlung fehlender Werte informieren möchte, kann eine elektronische Publikation des Rechenzentrums zu diesem Thema (Baltes-Götz 2008b) auf dem Webserver der Universität Trier von der Startseite (http://www.uni-trier.de/) ausgehend folgendermaßen finden:

Rechenzentrum > Studierende > EDV-Dokumentationen >
Statistik > Behandlung fehlender Werte in SPSS und Amos

7.5 Prüfung der KFA-Hypothese

Nun wollen wir die allgemeinpsychologische Kernhypothese unserer Studie prüfen, dass der Ärger über ein ungünstiges Ereignis durch die mentale Verfügbarkeit kontrafaktischer (also positiver) Alternativen gesteigert wird. Aufgrund der Ausreißer- und Verteilungsanalyse in Abschnitt 7.3.3 haben wir uns entschieden, statt des ursprünglich geplanten (parametrischen) t-Tests für abhängige Stichproben den verteilungsfreien **Vorzeichentest** zu verwenden. Suchen Sie die zu-

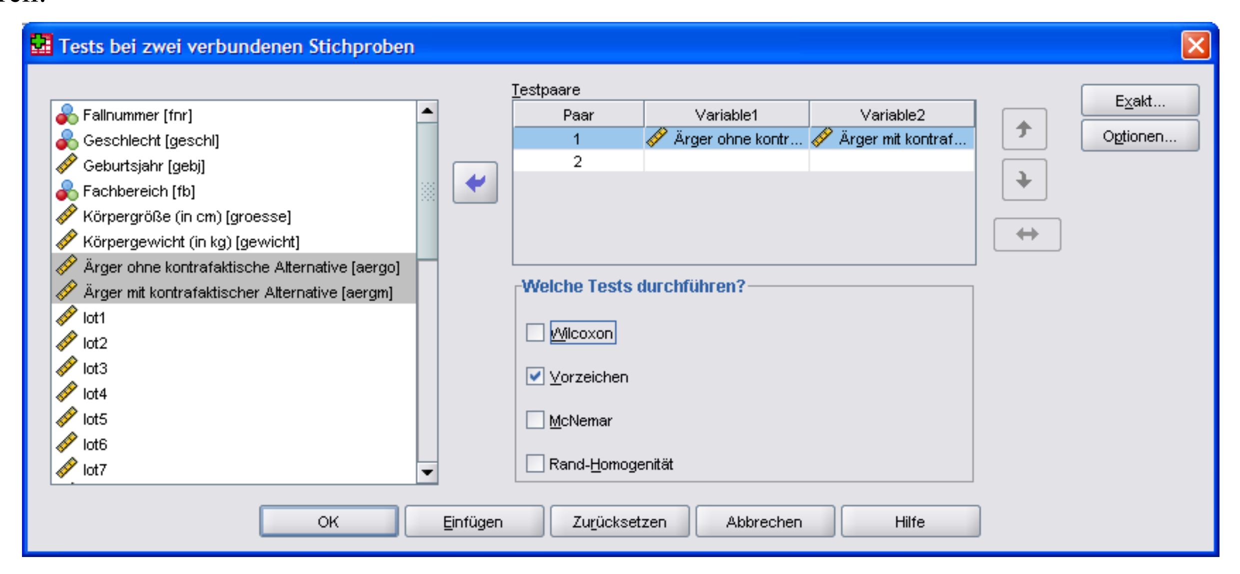
ständige Dialogbox zunächst über das **Analysieren**-Menü. Bei Misserfolg können Sie den Index des Hilfesystems benutzten, um einen Erklärungstext mit Wegweiser zur Menüposition zu finden. Steigen Sie ein mit

Hilfe > Themen > Index

und beginnen Sie dann, in das aktive Textfeld *Vorzeichentest* zu schreiben. Schon nach dem vierten Buchstaben wird der gesuchte Beitrag aufgelistet und kann per Doppelklick auf seinen Titel geöffnet werden. Im Hilfetext ist u.a. der Weg zur benötigten Dialogbox erklärt:

Analysieren > Nichtparametrische Tests > Zwei verbundene Stichproben

In der Dialogbox müssen Sie die beiden Variablen angeben und den gewünschten Test markieren:



Wir erhalten folgendes Ergebnis:

Häufigkeiten

		N
Ärger mit kontrafaktischer Alternative - Ärger ohne kontrafaktische Alternative	Negative Differenzen[a]	2
	Positive Differenzen[b]	26
	Bindungen[c]	3
	Gesamt	31

a. Ärger mit kontrafaktischer Alternative < Ärger ohne kontrafaktische Alternative

b. Ärger mit kontrafaktischer Alternative > Ärger ohne kontrafaktische Alternative

c. Ärger mit kontrafaktischer Alternative = Ärger ohne kontrafaktische Alternative

Statistik für Test[a]

	Ärger mit kontrafaktischer Alternative - Ärger ohne kontrafaktische Alternative
Z	-4,347
Asymptotische Signifikanz (2-seitig)	,000

a. Vorzeichentest

In unserer kleinen Stichprobe kann anstelle der per Voreinstellung gelieferten *asymptotischen* Überschreitungswahrscheinlichkeit auch die *exakte* ohne großen Zeitaufwand berechnet werden. Nach einem Mausklick auf den Schalter **Exakt** in obiger Dialogbox kann der exakte Test folgendermaßen angefordert werden:



Die unserer gerichteten Fragestellung entsprechende einseitige Überschreitungswahrscheinlichkeit ist deutlich kleiner als die konventionelle Grenze von 0,05. Damit kann die KFA-Nullhypothese (*Kein Ärgerzuwachs durch eine kontrafaktische Alternative*) zurückgewiesen werden:

Statistik für Test[a]

	Ärger mit kontrafaktischer Alternative - Ärger ohne kontrafaktische Alternative
Z	-4,347
Asymptotische Signifikanz (2-seitig)	,000
Exakte Signifikanz (2-seitig)	,000
Exakte Signifikanz (1-seitig)	,000
Punkt-Wahrscheinlichkeit	,000

a. Vorzeichentest

Nach Klärung der zentralen Hypothesen ist unser Projekt nun eigentlich abgeschlossen, aber es gibt noch viele SPSS-Optionen kennen zu lernen, und unsere Daten enthalten sicher auch noch einige interessante Details.

7.6 Übung

Für die Differenzvariable (GEWICHT - IDGEW) akzeptieren beide Normalverteilungstests die Nullhypothese:

Tests auf Normalverteilung

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
Gewicht - Idealgewicht	,092	31	,200*	,984	31	,905

a. Signifikanzkorrektur nach Lilliefors

*. Dies ist eine untere Grenze der echten Signifikanz.

Folglich darf mit den Variablen GEWICHT und IDGEW ein t-Test für verbundene Stichproben zu folgendem Testproblem durchgeführt werden (vgl. Abschnitt 7.2):

H_0: Das Realgewicht der Trierer Studierenden liegt im Mittel nicht unter dem Idealgewicht nach der Formel „Größe - 100“.

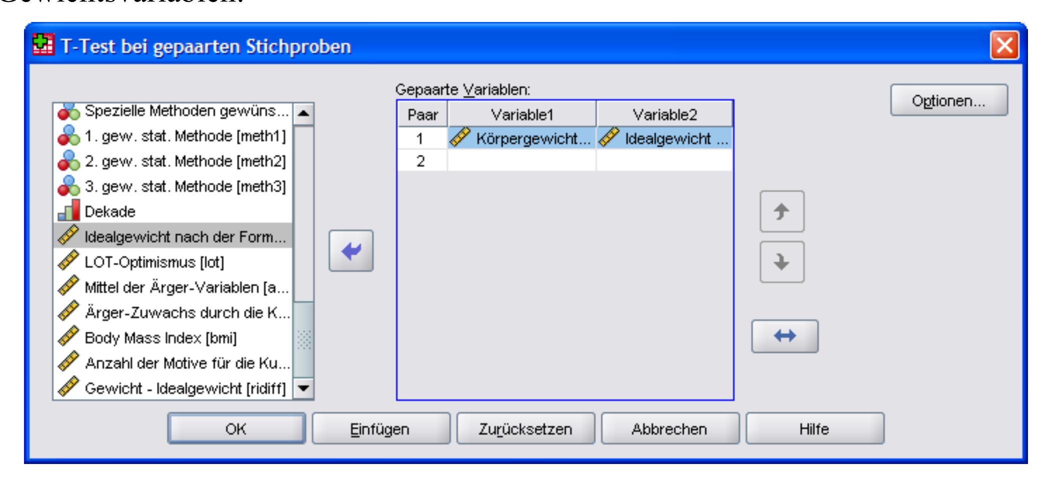
versus

H_1: Die Trierer Studierenden sind in Relation zur Idealgewichtsformel „Größe - 100“ im Mittel zu leicht.

Öffnen Sie mit dem Menübefehl

Analysieren > Mittelwerte vergleichen > T-Test bei gepaarten Stichproben

die zuständige Dialogbox, und bilden Sie über Markieren und Transportieren ein **Paar** aus den beiden Gewichtsvariablen:



Die Ergebnisse werden im nächsten Abschnitt vorgestellt.

7.7 Arbeiten mit dem Ausgabefenster (Teil II)

Oben wurde gelegentlich in didaktischer Nachlässigkeit der Begriff *Pivot-Tabelle* ohne Erläuterung verwendet. Unter dem *Pivotieren* einer Tabelle versteht SPSS u.a. die folgenden Operationen:

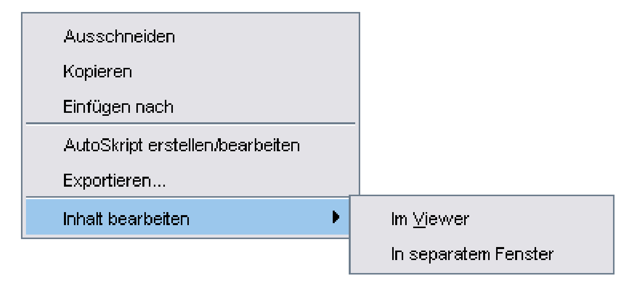
- Austauschen ihrer Zeilen-, Spalten- und Schichtendimensionen
- Änderung der Schachtelungsordnung
- Kategorien ausblenden

Neben diesen Pivot-Operationen bietet der Editor noch weitere Möglichkeiten zur Gestaltung von Ergebnistabellen.

7.7.1 Pivot-Editor starten

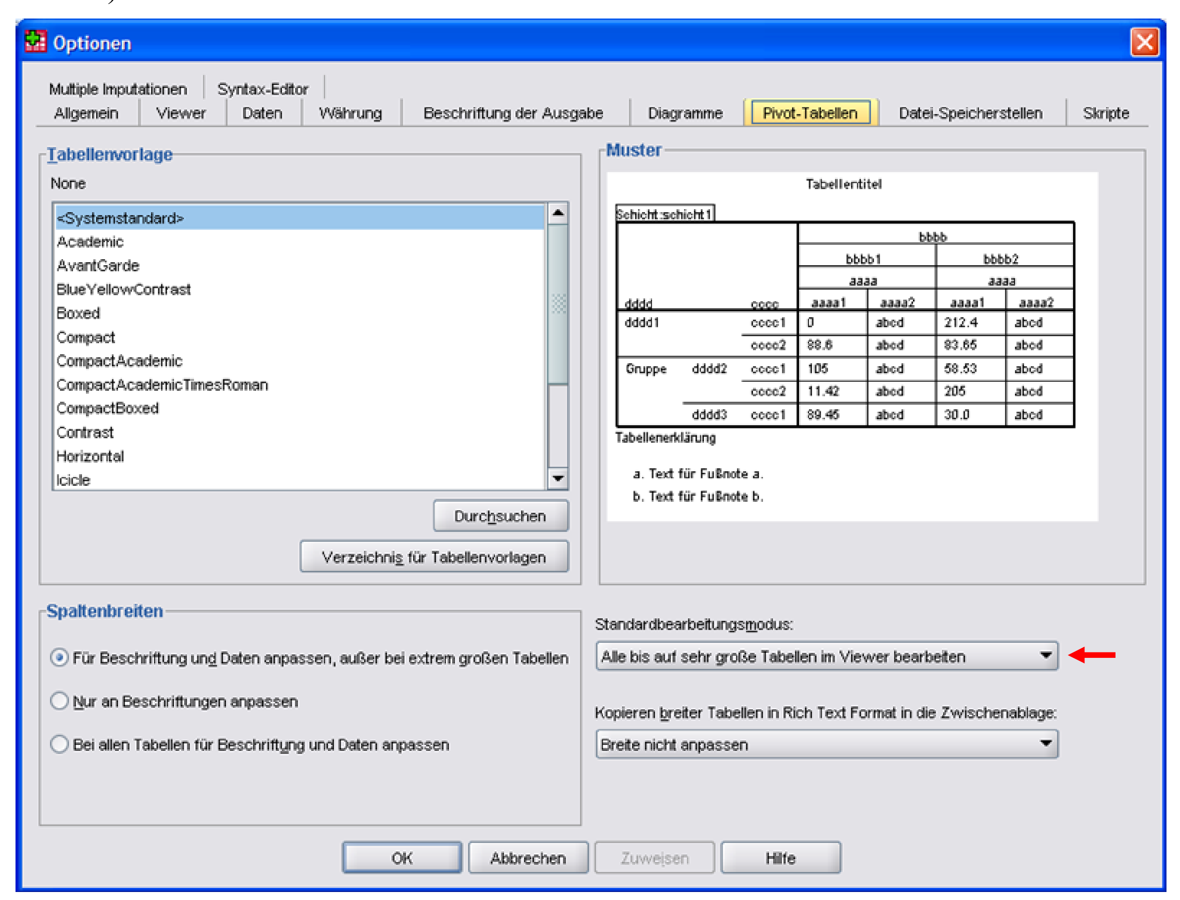
Um mit dem Editieren einer Tabelle zu beginnen, können Sie einen Doppelklick darauf setzen oder die Option **Inhalt bearbeiten** aus ihrem Kontextmenü wählen.

Bei der letztgenannten Methode bietet ein Untermenü

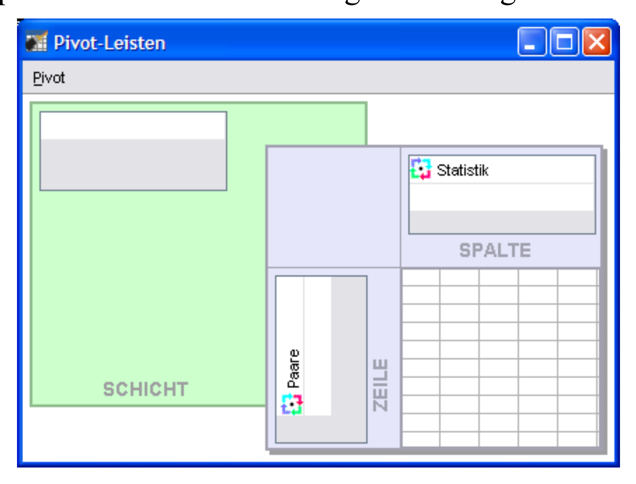


die Auswahl zwischen dem Bearbeiten innerhalb des **Viewers** und dem Öffnen eines **separaten Fensters**.

Ob ein Doppelklick zur Vor-Ort-Bearbeitung oder zum Öffnen eines separaten Fensters führt, hängt von der Größe der Tabelle und vom **Optionen**-Dialog ab (erreichbar über **Bearbeiten > Optionen**):



Für allgemeine Pivot-Operationen wird das das folgende Dialogfeld benötigt:



Es enthält je eine Ablagezone die Zeilen, Spalten und Schichten der Tabelle und je einen Eintrag mit Pivotsymbol für die dargestellten Tabellendimensionen. Sollten Sie das Dialogfeld vermissen, können Sie es mit dem folgenden Menübefehl aktivieren:

Pivot > Pivot-Leisten

Wir wollen als Beispiel die in obiger Übung von Ihnen erstellte Tabelle mit dem t-Test zum Vergleich von Real- und Idealgewicht betrachten:

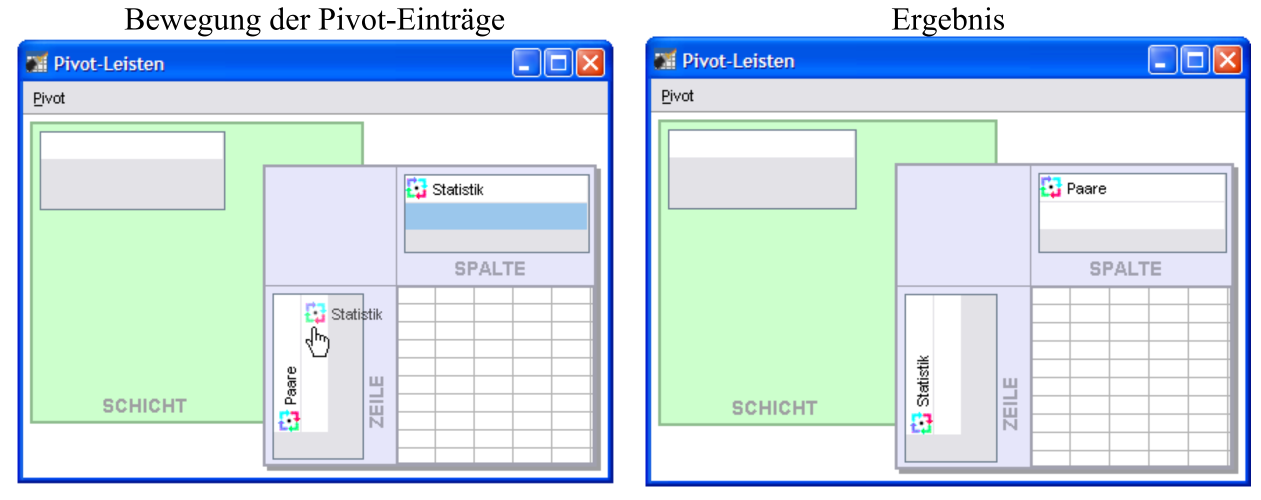
Test bei gepaarten Stichproben

		Gepaarte Differenzen					T	df	Sig. (2-seitig)
					95% Konfidenzintervall der Differenz				
		Mittelwert	Standardabweichung	Standardfehler des Mittelwertes	Untere	Obere			
Paaren 1	Körpergewicht (in kg) - Idealgewicht nach der Formel: Größe - 100	-9,323	6,188	1,111	-11,592	-7,053	-8,388	30	,000

Diese Tabelle enthält leider nur *eine* Schicht, so dass wir den Umgang mit Mehrschichttabellen nicht üben können. In den Zeilen der Tabelle wird die Dimension **Paare** dargestellt. Da wir nur ein einziges Variablenpaar untersucht haben, hat diese Dimension nur *eine* Kategorie. Die Spaltendimension **Statistik** sorgt mit ihren zahlreichen Kategorien für eine überbreite Tabelle, die schlecht auf ein DIN-A4-Blatt im Hochformat passt.

7.7.2 Dimensionen verschieben

Durch das Verschieben ihres Pivot-Eintrags kann man für eine Dimension neu festlegen, ob ihre Kategorien durch Spalten, Zeilen oder Schichten dargestellt werden sollen, z.B.:



Wenn in unserem Beispiel die beiden Pivot-Einträge bzw. Dimensionen ihre Plätze tauschen, benötigt die Tabelle in horizontaler Richtung deutlich weniger Platz:

Test bei gepaarten Stichproben

			Paaren 1
			Körpergewicht (in kg) - Idealgewicht nach der Formel: Größe - 100
Gepaarte Differenzen	Mittelwert		-9,323
	Standardabweichung		6,188
	Standardfehler des Mittelwertes		1,111
	95% Konfidenzintervall der Differenz	Untere	-11,592
		Obere	-7,053
T			-8,388
df			30
Sig. (2-seitig)			,000

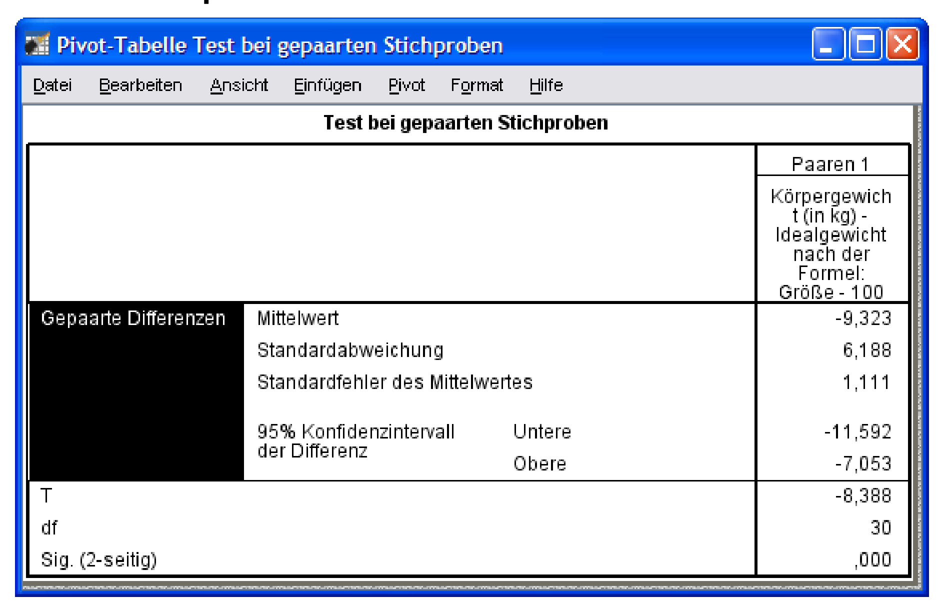
Wenn es lediglich um das Vertauschen von Zeilen und Spalten einer Tabelle geht, kann man an Stelle des flexiblen Pivot-Werkzeugs auch den Menübefehl

Pivot > Zeilen und Spalten vertauschen

verwenden.

7.7.3 Gruppierungen

Kategorien einer Dimension können zu einer Gruppe zusammengefasst und durch eine etikettierende Zelle hervorgehoben sein. In unserem Beispiel zeigt sich bei der Statistikdimension eine Gruppe mit dem Etikett **Gepaarte Differenzen**:



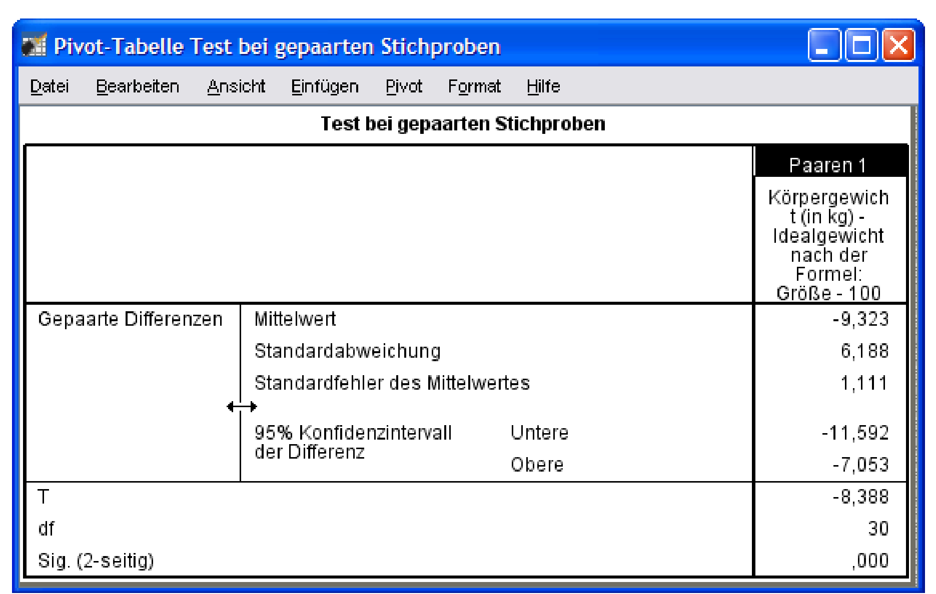
Test bei gepaarten Stichproben

			Paaren 1
			Körpergewicht (in kg) - Idealgewicht nach der Formel: Größe - 100
Gepaarte Differenzen	Mittelwert		-9,323
	Standardabweichung		6,188
	Standardfehler des Mittelwertes		1,111
	95% Konfidenzintervall der Differenz	Untere	-11,592
		Obere	-7,053
T			-8,388
df			30
Sig. (2-seitig)			,000

In SPSS 15 war es noch möglich, überflüssige Gruppierung folgendermaßen zu beseitigen:

- Rechtsklick auf das Kategorienetikett
- Aus dem Kontextmenü wählen: **Gruppierung aufheben**

Leider hat sich in SPSS 17 ein Fehler eingeschlichen, so dass die benötigte Kontextmenüoption nicht mehr nutzbar ist. Allerdings lässt sich in SPSS 17 eine Gruppenzelle zu *Zeilen*kategorien doch entfernen, indem man die Breite der Zelle durch Verschieben des rechten Rands auf Null bringt:



Test bei gepaarten Stichproben

			Paaren 1
			Körpergewicht (in kg) - Idealgewicht nach der Formel: Größe - 100
Gepaarte Differenzen	Mittelwert		-9,323
	Standardabweichung		6,188
	Standardfehler des Mittelwertes		1,111
	95% Konfidenzintervall der Differenz	Untere	-11,592
		Obere	-7,053
T			-8,388
df			30
Sig. (2-seitig)			,000

Um denselben Trick bei der **Paare**-Dimension auf die mit **Paaren 1** beschriftete Gruppenzelle anwenden zu können, bringt man die Dimension vorübergehend in den Zeilenbereich. Gegen die verbliebene Gruppe mit den Schranken zum **95% Konfidenzintervall der Differenz** ist nichts einzuwenden:

Test bei gepaarten Stichproben

		Körpergewicht (in kg) - Idealgewicht nach der Formel: Größe - 100
Mittelwert		-9,323
Standardabweichung		6,188
Standardfehler des Mittelwertes		1,111
95% Konfidenzintervall der Differenz	Untere	-11,592
	Obere	-7,053
T		-8,388
df		30
Sig. (2-seitig)		,000

Das trickreiche Entfernen überflüssiger Gruppierungen hat leider schwer vorhersehbare Effekte auf die Gitterlinien der Tabelle.

Wenn Sie mehrere Kategorien einer Dimension zu einer Gruppe zusammenfassen wollen, können Sie folgendermaßen vorgehen:

- Alle Kategorien markieren
- Kontextmenü zu einer markierten Kategorie öffnen und Option **Gruppe** wählen
- Beschriftung der Gruppenzelle nach Doppelklick anpassen

In der folgenden Version unserer Tabelle wurde eine Gruppe mit den drei Kategorien zum t-Test gebildet:

Test bei gepaarten Stichproben

			Körpergewicht (in kg) - Idealgewicht nach der Formel: Größe - 100
Mittelwert			-9,323
Standardabweichung			6,188
Standardfehler des Mittelwertes			1,111
95% Konfidenzintervall der Differenz	Untere		-11,592
	Obere		-7,053
Signifikanztest		T	-8,388
		df	30
		Sig. (2-seitig)	,000

Außerdem wurde das Gruppenetikett vertikal zentriert über das Registerblatt **Ausrichtung und Ränder** der per Kontextmenü erreichbaren Dialogbox mit den **Zelleneigenschaften**. Leider nervt erneut das unberechenbare Auftauchen und Verschwinden von Gitterlinien.

Selbst definierte Gruppierungen lassen sich über das Kontextmenü-Item **Gruppierung aufheben** wieder entfernen.

7.7.4 Kategorien aus- und einblenden

Wenn eine SPSS-Tabelle zu ausführlich erscheint, können Kategorien einer Dimension ausgeblendet werden. In unserem Beispiel wollen wir bei der Statistikdimension auf den Standardfehler des Mittelwerts verzichten:

Test bei gepaarten Stichproben

			Körpergewicht (in kg) - Idealgewicht nach der Formel: Größe - 100
Mittelwert			-9,323
Standardabweichung			6,188
95% Konfidenzintervall der Differenz	Untere		-11,592
	Obere		-7,053
Signifikanztest		T	-8,388
		df	30
		Sig. (2-seitig)	,000

Gehen Sie beim Ausblenden einer Kategorie folgendermaßen vor:

- Bei gedrückter Tastenkombination **Strg+Alt** einen (linken) Mausklick auf das Kategorienetikett setzen, um die Kategorie komplett zu markieren
- Rechtsklick auf das Kategorienetikett
- Aus dem Kontextmenü wählen: **Kategorie ausblenden**

In *Spalten* untergebrachte Kategorien kann man auch auf intuitive Weise eliminieren:

- linker Mausklick auf den rechten Spaltenrand, Maustaste gedrückt halten
- Spaltenbreite durch Verschieben der Maus reduzieren, bis die Quick-Info **Ausblenden** erscheint:

Gepaarte Differenzen			
Standardabweichung	Standardfehler des Mittelwerte	95% Konfidenzintervall der Differenz Ausblenden	 Obere
6,188	1,111	-11,592	-7,053

- Maustaste loslassen

Zum *Einblenden* von vorher abgeschalteten Kategorien kenne ich (neben **Bearbeiten > Rückgängig**) nur die global wirksame Methode:

Ansicht > Alles einblenden

Nach diesem Befehl können Tabellenbestandteile auftauchen (z.B. Dimensionsbeschriftungen), die (je nach verwendeter Vorlage) bei *neuen* Tabellen nicht eingeschaltet sind.

7.7.5 Zellen modifizieren

Text editieren

Bei aktivem Pivot-Editor können Sie nach einem Doppelklick auf eine Zelle den enthaltenen Text beliebig ändern. In unserem Beispiel sollte der Titel etwas informativer und die Beschriftung der rechten Spalte etwas sparsamer werden. Außerdem sollten wir konsistent mit der Hypothesenformulierung (siehe Abschnitt 7.6) die *einseitige* Überschreitungswahrscheinlichkeit angeben. Im aktuellen Beispiel wirkt sich allerdings das Halbieren der betragsmäßig sehr kleinen Zahl auf den ersten drei Dezimalstellen nicht aus:

t-Test zur Differenz von Real- und Idealgewicht

			Real - Idealgewicht
Mittelwert			-9,323
Standardabweichung			6,188
Standardfehler des Mittelwertes			1,111
95% Konfidenzintervall der Differenz	Untere		-11,592
	Obere		-7,053
Signifikanztest		T	-8,388
		df	30
		Sig. (1-seitig)	,000

Zellen zur weiteren Bearbeitung markieren

Mit dem Menübefehl **Bearbeiten > Auswählen** lassen sich Tabellenbestandteile (z.B. Tabellenkorpus, Datenzellen) zur weiteren Bearbeitung markieren. Außerdem stehen die Windows-üblichen Markierungsmethoden per Maus und Tastatur zur Verfügung.

Zelleneigenschaften

Über die per Kontextmenü erreichbare Dialogbox mit den **Zelleneigenschaften** können zahlreiche Attribute der markierten Zellen beeinflusst werden:

- Schriftart, Text- und Hintergrundfarbe
- Zahlenformate, Anzahl der Dezimalstellen
- Ausrichtung der Zellinhalte
- Randabstände der Zellinhalte

Um die Anzahl der Dezimalstellen anzupassen, kann man so vorgehen:

- Alle betroffenen Zellen markieren
- Menübefehl **Format > Zelleneigenschaften** wählen
- Auf der **Formatwert**-Registerkarte die gewünschte Anzahl der **Dezimalstellen** eintragen

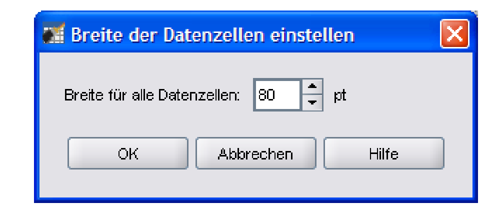
d) Spaltenbreite

Wenn sich der Mauszeiger über dem rechten Rand einer Spalte befindet, ändert er seine Form zu einem doppelseitigen Pfeil ⟷. Jetzt können Sie durch Klicken und Ziehen bei gedrückter linker Maustaste die Spaltengrenze verschieben und somit die Spaltenbreite ändern.

Über den Menübefehl

Format > Breite der Datenzellen

lässt sich die Breite sämtlicher Datenzellen einer Tabelle numerisch spezifizieren, z.B.:



Nach missratenen Gestaltungsbemühungen bringt eventuell der Menübefehl

Format > Automatisch anpassen

wieder ein akzeptables Ergebnis zu Stande.

7.7.6 Tabellenvorlagen

Für eine Pivot-Tabelle kann nach dem Menübefehl

Format > Tabellenvorlagen

das Design einer Tabellenvorlage übernommen werden. So sieht unser Beispiel nach Anwendung der Vorlage **Academic** aus:

t-Test zur Differenz von Real- und Idealgewicht

			Real - Idealgewicht
Mittelwert			-9,32
Standardabweichung			6,19
Standardfehler des Mittelwertes			1,11
95% Konfidenzintervall der Differenz	Untere		-11,59
	Obere		-7,05
Signifikanztest		T	-8,39
		df	30
		Sig. (1-seitig)	0,00

An das sinnlose Auftauchen und Verschwinden von Gitterlinien haben wir uns mittlerweile gewohnt. Solche Fehler können in der Regel nach dem Zwischenablagen-Transfer zu einem Textverarbeitungsprogramm mit den dortigen Mitteln behoben werden.

8 Gruppenvergleiche

In diesem Abschnitt interessieren wir uns für Geschlechtsunterschiede beim Body Mass Index und führen mit unseren Variablen GESCHL und BMI einen t-Test für unabhängige Stichproben zum folgenden Hypothesenpaar durch:

H_0: Bei Frauen ist der BMI-Mittelwert mindestens genauso groß wie bei Männern.

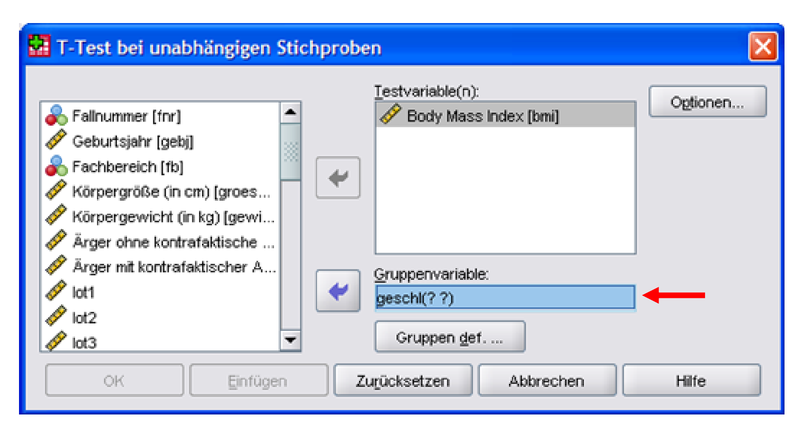
versus

H_1: Bei Frauen ist der BMI-Mittelwert niedriger als bei Männern.

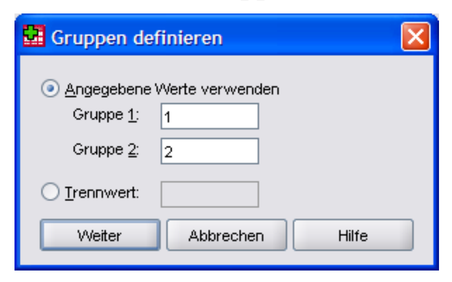
Fordern Sie mit folgendem Menübefehl die zugehörige Dialogbox an:

Analysieren > Mittelwerte vergleichen > T-Test bei unabhängigen Stichproben

Transportieren Sie BMI in die Liste der **Testvariable(n)** und GESCHL in das Feld **Gruppenvariable**:



Über den Schalter **Gruppen def.** erreicht man die folgende Dialogbox, um die beiden zu vergleichenden Gruppen über ihre Werte bei der Gruppenvariablen festzulegen:



In unserem Fall sind nur zwei Gruppen vorhanden, die folglich beide teilnehmen sollen.

Wir erhalten folgende Ergebnisse:

Gruppenstatistiken

	Geschlecht	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
Body Mass Index	Frau	25	20,7488	1,89347	,37869
	Mann	6	22,8078	2,17495	,88792

Bei den Männern fällt der BMI-Mittelwert im H_1-Sinn um ca. 2 Punkte höher aus.

Zunächst ist die Frage zu klären, welche der beiden angebotenen t-Test – Varianten *(mit* bzw. *ohne* Voraussetzung der Varianzhomogenität) zu verwenden ist. Als Entscheidungshilfe berechnet SPSS den **Levene-Test der Varianzhomogenität**, der in unserem Fall durch eine empirische Überschreitungswahrscheinlichkeit von 0,94 (> 0,05) seine Nullhypothese gleicher Varianzen akzeptiert.

Test bei unabhängigen Stichproben

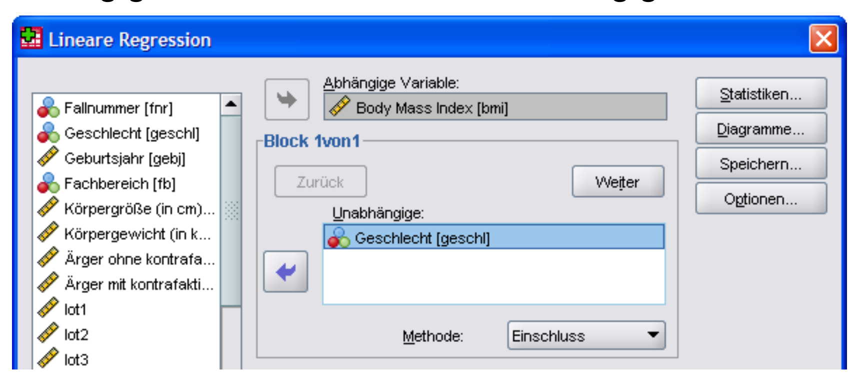
		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
									95% Konfidenzintervall der Differenz	
		F	Signifikanz	T	df	Sig.(2-seitig)	Mittlere Differenz	Standardfehler der Differenz	Untere	Obere
Body Mass Index	Varianzen sind gleich	,006	,940	-2,329	29	,027	-2,05895	,88417	-3,86727	-,25062
	Varianzen sind nicht gleich			-2,133	6,937	,071	-2,05895	,96530	-4,34576	,22787

Der somit verwendbare klassische t-Test *mit* vorausgesetzter Varianzhomogenität ermittelt eine Überschreitungswahrscheinlichkeit unterhalb der kritischen Grenze von 0,05, so dass die Nullhypothese zu verwerfen ist, sofern die Voraussetzungen des Test hinreichend erfüllt sind. Weil ein einseitiges (gerichtetes) Testproblem vorliegt, wäre auch der bei einem signifikanten Levene-Ergebnis zu verwendende t-Test *ohne* Varianzhomogenitätsannahme zur selben Entscheidung gekommen.

Nachdem die Varianzhomogenität der Residuen geklärt ist, und deren Unabhängigkeit angenommen werden darf, bleibt von den Voraussetzungen der Analyse noch die Normalität der Residuen zu untersuchen. Um die Verteilung der Residuen mit geringem technischem Aufwand per Histogramm beurteilen zu können, führen wir den t-Test für unabhängige Stichproben mit der Prozedur für die lineare Regression erneut durch. Diese Prozedur beherrscht als Spezialfall auch den klassischen t-Test (mit angenommener Varianzhomogenität) und bietet generell die Ausgabe eines Histogramms zu den Residuen an. Nach dem Menübefehl

Analysieren > Regression > Linear

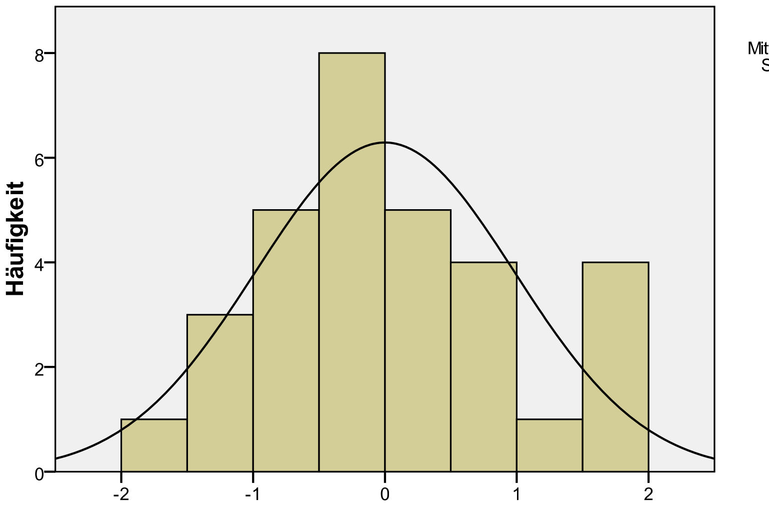
wählen wir die **anhängige Variable** BMI und die **unabhängige Variable** GESCHL:



In der Subdialogbox **Diagramme** fordern wir ein **Histogramm** für die standardisierten Residuen an:



Das resultierende Histogramm gibt keinen Anlass zur Sorge bzgl. der Normalverteilungsannahme:



Die über eine **explorative Datenanalyse** (siehe Abschnitt 7.3.2) für die abgespeicherten standardisierten Residuen (siehe Abschnitt 7.4.1) durchgeführten Signifikanztests zur Normalitäts-Nullhypothese bestätigen den visuellen Eindruck:

Tests auf Normalverteilung

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
Standardized Residual	,133	31	,171	,960	31	,291

a. Signifikanzkorrektur nach Lilliefors

In der Koeffiziententabelle der linearen Regression findet sich erwartungsgemäß das t-Testergebnis wieder, dessen Interpretierbarkeit mittlerweile bestätigt ist:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		Regressions-koeffizient B	Standardfehler	Beta	T	Sig.
1	(Konstante)	18,690	1,112		16,813	,000
	Geschlecht	2,059	,884	,397	2,329	,027

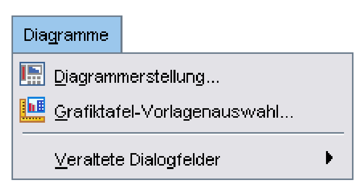
a. Abhängige Variable: Body Mass Index

Es stellt übrigens *kein* Problem dar, dass die beiden Stichproben verschieden groß sind. Man sollte bei der Untersuchungsplanung nach Möglichkeit für gleich große Teilstichproben sorgen, weil bei dieser Aufteilung eine optimale Teststärke resultiert und außerdem eine gewisse Robustheit gegen Verletzungen der Varianzhomogenität. Sind aber Daten mit ungleicher Aufteilung vorhanden, spricht nichts gegen ihre Verwendung, zumal beim t-Test für unabhängige Stichproben die Voraussetzung der Varianzhomogenität vermieden werden kann.

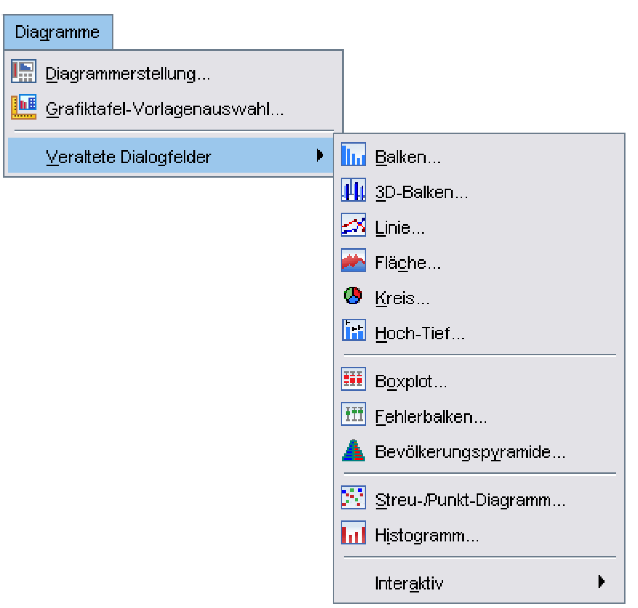
9 Graphische Datenanalyse

Wir haben schon einige graphische Darstellungsmöglichkeiten kennen gelernt, die im Rahmen von Statistikprozeduren angeboten werden (z.B. Histogramm, Boxplot). In diesem Abschnitt arbeiten wir erstmals mit dem **Diagramme**-Menü und vor allem mit dem Diagramm-Editor zur individuellen Nachbearbeitung von Diagrammen.

SPSS-Einsteiger werden vermutlich durch das **Diagramme**-Menü leicht irritiert, weil hier gleich *drei* Zugänge angeboten werden:



Über **veraltete Dialogfelder**



(verknüpft mit dem SPSS-Kommando GRAPH) oder mit dem Dialog **Diagrammerstellung** (verknüpft mit dem SPSS-Kommando GGRAPH und der *Graphics Production Language* (GPL)) entstehen Diagramme, die anschließend mit dem Diagramm-Editor modifiziert werden können. Seit SPSS 17 ist über die **Grafiktafel-Vorlagenauswahl** ein weiterer Assistent zur Diagrammerstellung verfügbar. Obwohl im Hintergrund ebenfalls das SPSS-Kommando GGRAPH beteiligt ist, gehören die Diagramme offenbar zu einer neuen Ausgabekategorie und werden mit dem neuen Grafiktafel-Editor bearbeitet. Ein Blick in die SPSS-Handbücher lässt den Schluss zu, dass der Hersteller von den drei möglichen Einstiegen in die Diagrammproduktion derzeit wohl die **Diagrammerstellung** empfiehlt.

Von den zahlreich angebotenen Graphiktypen können aus Zeitgründen nur wenige Beispiele behandelt werden. Im aktuellen Abschnitt 9 wird das Streudiagramm vorgestellt, in Abschnitt 11.2 kommt ein Balkendiagramm zum Einsatz.

9.1 Streudiagramm anfordern

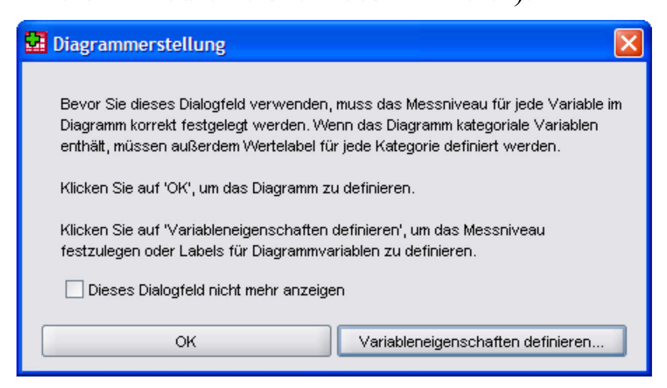
Um die empirische Regression von Gewicht auf Größe und Geschlecht betrachten zu können, fordern wir ein Streudiagramm mit diesen Variablen an. Dies tun wir (mit grundsätzlich identischem Ergebnis) sowohl mit der Dialogbox **Diagrammerstellung** als auch mit einem **veralteten Dialogfeld**.

9.1.1 Diagrammerstellung

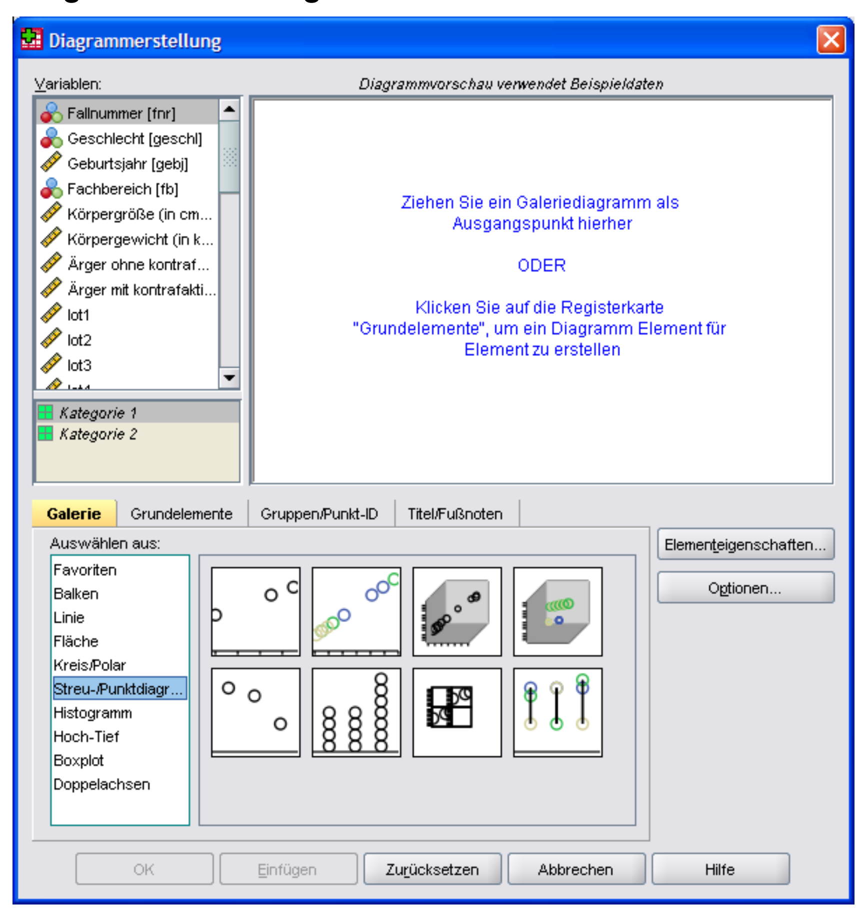
Nach dem Menübefehl

Diagramme > Diagrammerstellung

informiert SPSS zunächst darüber, dass bei allen Variablen korrekt deklarierte Messniveaus und bei kategorialen (ordinalen oder nominalen) Variablen außerdem Wertelabels benötigt werden (zur Deklaration von Variablenattributen siehe Abschnitt 3.2.2):



Das Dialogfeld **Diagrammerstellung**



unterstützt zwei Vorgehensweisen zur Definition eines neuen Diagramms:

- Graphiktyp aus der **Galerie** als Ausgangspunkt wählen und individuell gestalten
- Graphik aus **Grundelementen** (z.B. Achsensystem, Linie) aufbauen

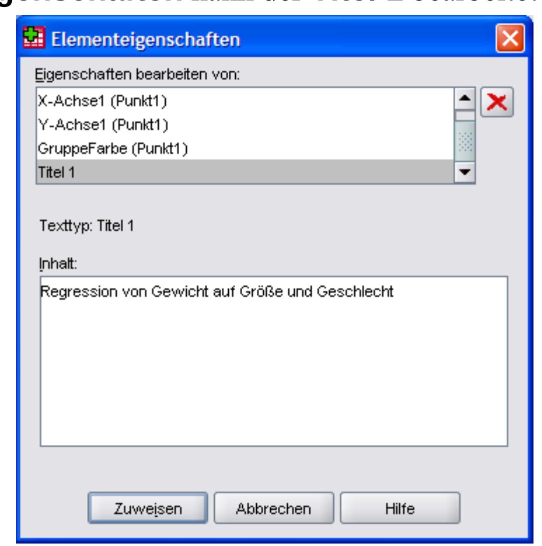
Wir wählen den vom Hersteller empfohlenen *ersten* Weg:

- Klicken Sie auf die Registerkarte **Galerie**, und wählen Sie den Typ **Streu-/Punkt-diagramm**.

- Ziehen Sie das Symbol zum **gruppierten Streudiagramm** auf die Diagrammvorschau (Entwurfszone) über den Diagrammtypen.
- In der Diagrammvorschau erscheint ein Achsensystem mit Ablageflächen für
  - eine X-Achsen-Variable
  - eine Y-Achsen-Variable
  - eine Gruppierungsvariable (Beschriftung: **Farbe festlegen**)

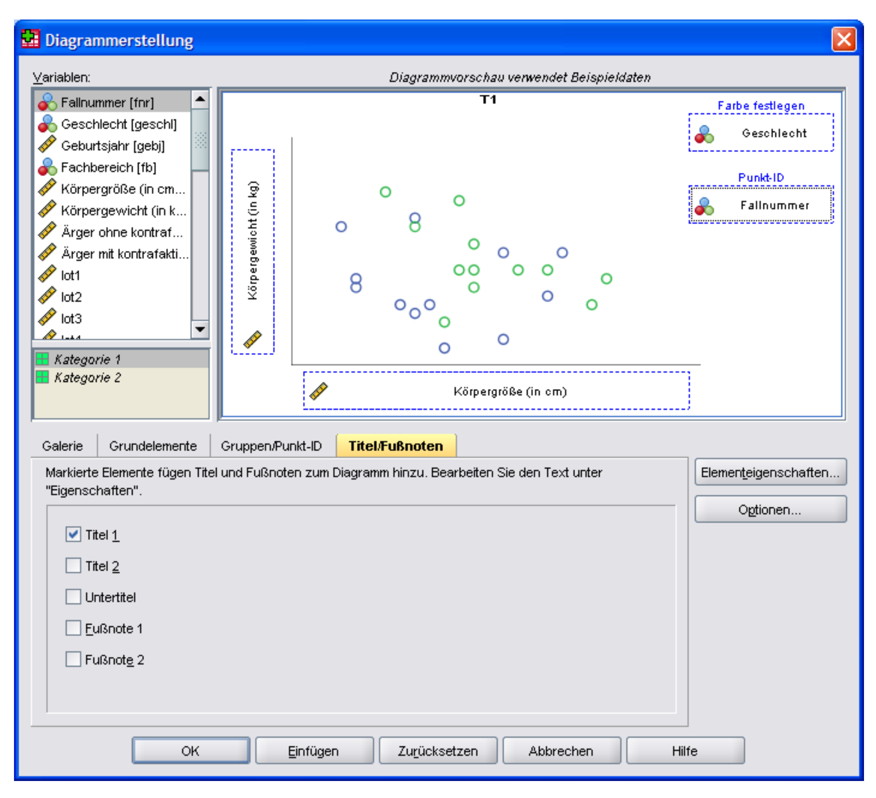
  Außerdem erscheint die zusätzliche Dialogbox **Elementeigenschaften**.
- Bringen Sie nun die drei Variablen GROESSE, GEWICHT und GESCHL in Position:
  - Ziehen Sie aus der Liste in der linken oberen Ecke die Variable GROESSE auf die X-Achsen-Ablagefläche.
  - Ziehen Sie die Variable GEWICHT auf die Y-Achsen-Ablagefläche.
  - Ziehen Sie die Variable GESCHL auf die Gruppierungs-Ablagefläche.
    So erhält man für weibliche und männliche Datenpunkte verschiedene Markierungen und kann ggf. geschlechtsbedingte Unterschiede bei der Regression von Gewicht auf Größe erkennen.

  Zur Illustration werden künstliche Datenpunkte angezeigt.
- Gehen Sie folgendermaßen vor, um die Variable FNR zur Fallbeschriftung nutzen zu können:
  - Klicken Sie auf die Registerkarte **Gruppen/Punkt-ID**, und markieren Sie das Kontrollkästchen **Punkt-ID-Beschriftung**.
  - Daraufhin erscheint die neue Ablagefläche **Punktbeschriftungsvariable** in der Diagrammvorschau. Ziehen Sie die Variable FNR dorthin.
- Legen Sie einen Titel für die Graphik fest:
  - Klicken Sie auf die Registerkarte **Titel/Fußnoten**, und markieren Sie das Kontrollkästchen **Titel 1**.
  - Daraufhin erscheint in der Diagrammvorschau der Platzhalter **T1**, und in der Dialogbox **Elementeigenschaften** kann der **Titel 1** bearbeitet werden, z.B.



    Tragen Sie einen Text ein, und quittieren Sie mit einem Mausklick auf den Schalter **Zuweisen**.

Nun sollte die Dialogbox **Diagrammerstellung** ungefähr folgendes Bild zeigen:



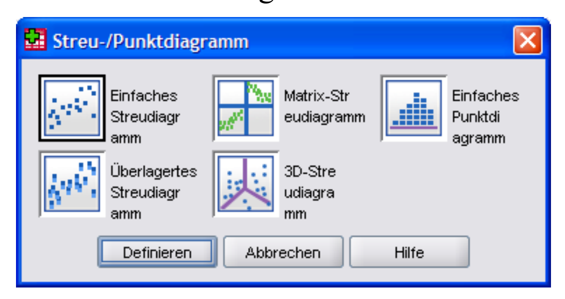
Nach einem Klick auf den Schalter **OK** wird die Graphik erstellt. Das Ergebnis ist in Abschnitt 9.2 zu sehen.

9.1.2 Dialogbox Einfaches Streudiagramm

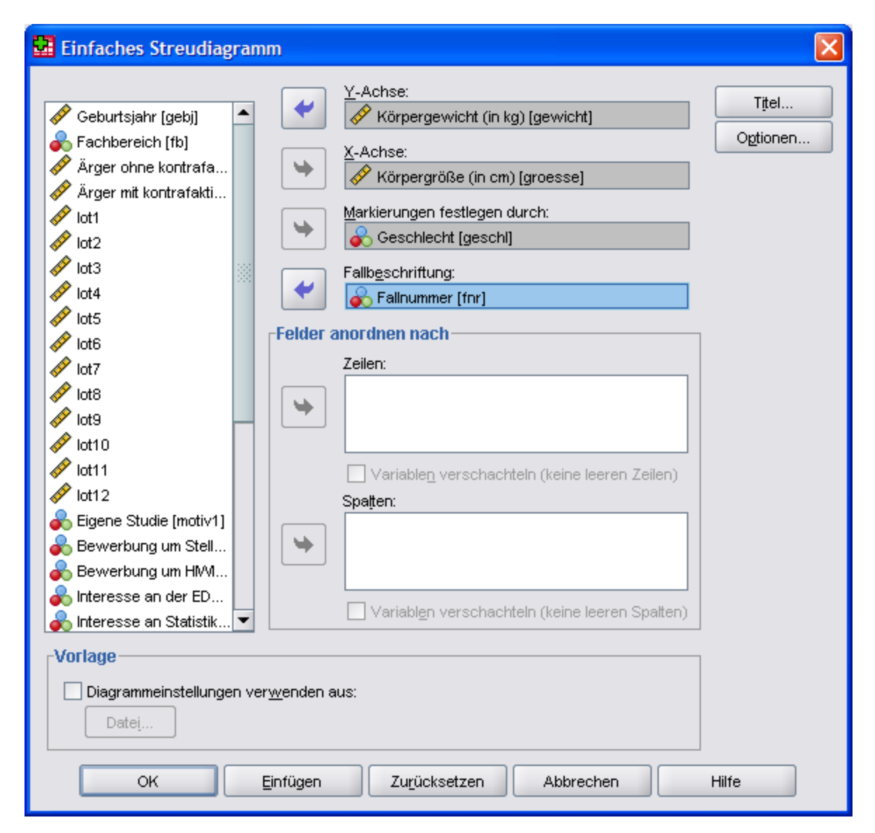
Wer sich mit der **Diagrammerstellung** noch nicht anfreunden kann, hat in der SPSS-Version 17 auch noch die **veralteten Dialogfelder** zur Verfügung, z.B. zum Erstellen eines Streudiagramms:

Diagramme > Veraltete Dialogfelder > Streu-/Punkt-Diagramm

In der nach diesem Menübefehl erscheinenden Palette akzeptieren wir für das Streudiagramm mit Gewicht, Größe und Geschlecht die voreingestellte einfache Variante



und wechseln per Mausklick auf den Schalter **Definieren** zur Dialogbox **Einfaches Streudiagramm**, wo die beteiligten Variablen per Drag & Drop oder Transportschalter ihre Rollen erhalten:



Durch die Verwendung von GESCHL als **Markierungesvariable** werden weibliche und männliche Datenpunkte verschieden dargestellt, so dass geschlechtsbedingte Unterschiede bei der Regression von Gewicht auf Größe ggf. sichtbar werden.

Die Variable FNR soll später im **Datenbeschriftungsmodus** verwendet werden (siehe Abschnitt 9.2).

Nach einem Mausklick auf den Schalter **Titel** tragen wir eine Titelzeile ein:



Quittieren Sie die Subdialogbox mit **Weiter** und die Hauptdialogbox mit **OK**, um die Graphik zu erstellen.

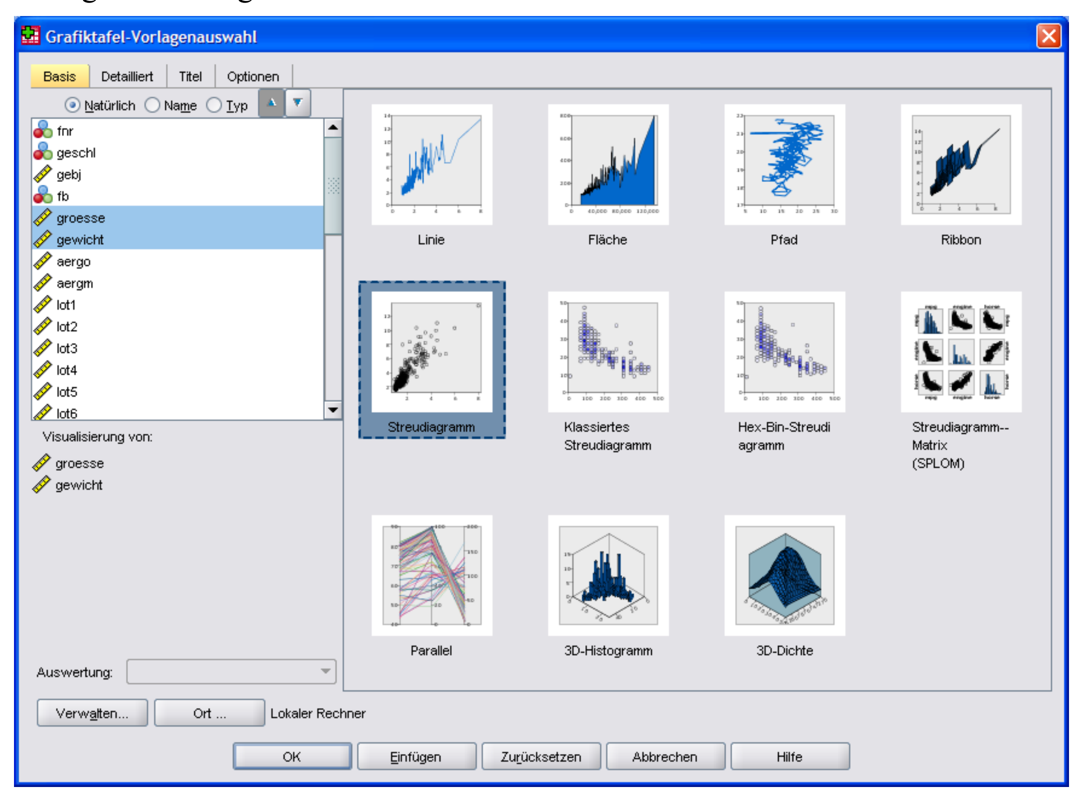
Das Erstellen eines einfachen Streudiagramms gelingt mit den veralteten Dialogfeldern ebenso gut wie mit der Diagrammerstellung. Zudem wird sich in Abschnitt 9.2 zeigen, dass die konventionell erstellten Streudiagramme bei der Modifikation im Graphik-Editor weniger Probleme machen. Generell lassen sich allerdings per Diagrammerstellung mehr Darstellungswünsche realisieren als mit den veralteten Dialogfeldern.

9.1.3 Grafiktafel-Vorlagenauswahl

Nach dem Menübefehl

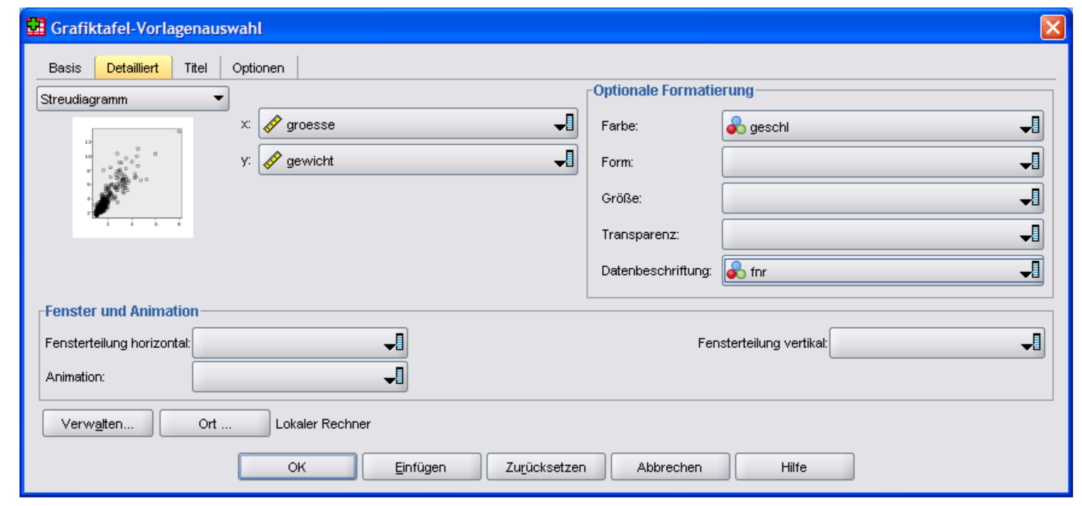
Diagramme > Grafiktafel-Vorlagenauswahl

bietet der folgende Dialog



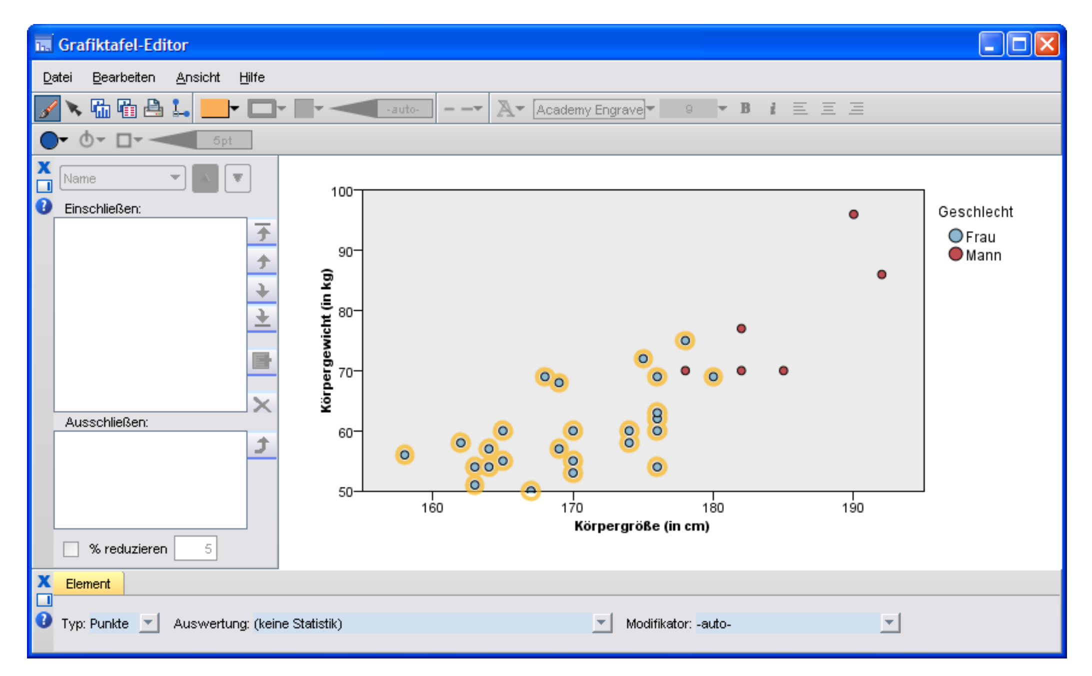
analog zur Diagrammerstellung einen Basis und einen Detail-Modus:

- Auf der Registerkarte **Basis** erscheinen nach Auswahl von Variablen und **Auswertungs**-Statistik Visualisierungsvorschläge zur direkten Übernahme durch Markieren und Quittieren.
- Auf der Registerkarte **Detailliert**



geschieht der Diagrammentwurf wie in einem traditionellen (veralteten) Dialog. Man kann selbstverständlich nach Wahl einer **Basis**-Visualisierung noch **Detail**-Angaben machen, z.B. eine Markierungs- und/oder eine Datenbeschriftungsvariable wählen.

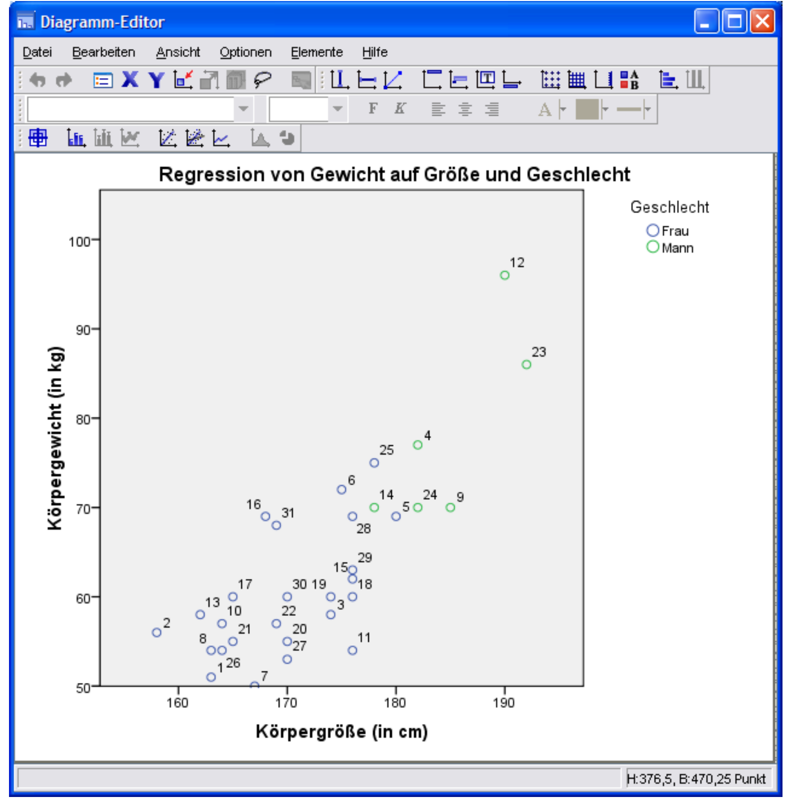
Öffnet man im Ausgabefenster ein per Grafiktafel-Vorlagenauswahl erstelltes Diagramm, erscheint der Grafiktafel-Editor, z.B.:



Er bietet im Vergleich zum Diagramm-Editor (siehe Abschnitt 9.2) sehr eingeschränkte Möglichkeiten und erlaubt z.B. nicht das Ergänzen einer Regressionsgeraden. Wir werden die Graphiktafel-Technik in diesem Kurs nicht mehr benutzen.

9.2 Streudiagramm per Diagramm-Editor modifizieren

Wenn Sie im Ausgabefenster einen Doppelklick auf eine per **Diagramerstellung** oder über ein **veraltetes Dialogfeld** erstellte Graphik setzen, wird sie im Diagramm-Editor geöffnet:



Anschließend werden am Beispiel des Streudiagramms einige allgemeine Bedienungsmöglichkeiten des Diagramm-Editors vorgestellt. Deren Effekte lassen sich über die Schalter ↶ ↷ (mehrstufig) rückgängig machen bzw. wiederherstellen.

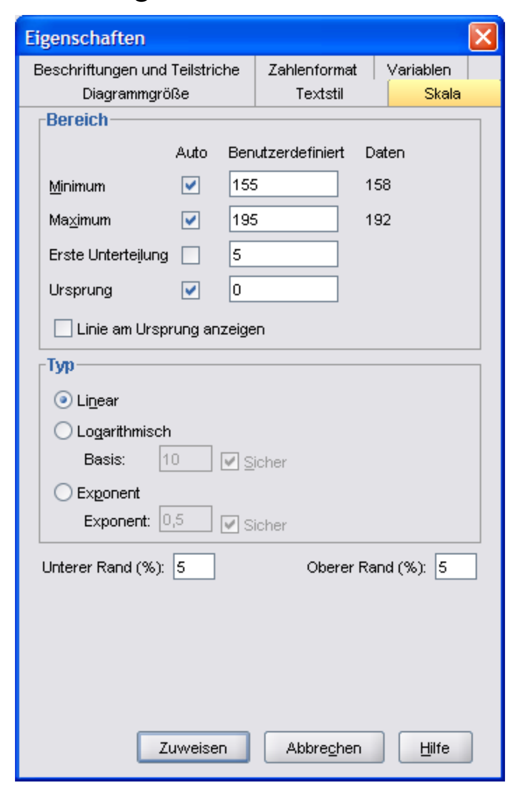
Vorweg soll schon verraten werden, wie die Datenbeschriftungen abzuschalten sind, die SPSS übereifrig eingetragen hat, weil wir bei der Diagrammerstellung (siehe Abschnitt 9.1.1) FNR zur **Punktbeschriftungsvariablen** ernannt haben:

- Mausklick auf den Schalter
- oder Menübefehl **Elemente > Datenbeschriftungen ausblenden**

Bei Verwendung des veralteten Dialogfelds (vgl. Abschnitt 9.1.2) sind die Datenbeschriftungen trotz analoger Vorgehensweise per Voreinstellung ausgeblendet.

9.2.1 Eigenschaftsfenster

Zum aktuell im Diagramm-Editor markierten Objekt bzw. zur markierten Objektgruppe (erkennbar an einer Umrahmung) bietet das Eigenschaftsfenster



auf jeweils dynamisch erstellten Registerkarten alle modifizierbaren Attribute. Bei Bedarf kann es per Doppelklick auf ein zu gestaltendes Objekt, über das Symbol , mit der Tastenkombination **Strg+T** oder mit den Menübefehl

Bearbeiten > Eigenschaften

aktiviert werden.

Wer im Beispiel X-Achsenteilstrichwerte im Abstand von 5 cm wünscht, kann so vorgehen:

- X-Achsenteilstrichwerte per Mausklick auf einen Wert markieren



- im Eigenschaftsfenster die Registerkarte **Skala** wählen (siehe oben)
- bei der **ersten Unterteilung** den **benutzerdefinierten** Wert 5 eintragen
- **Zuweisen**

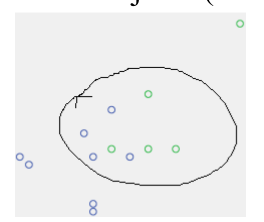
9.2.2 Markieren von gruppierten Objekten

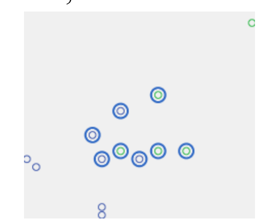
Sind gruppierte Objekte vorhanden (z.B. die Datenpunkte für Frauen bzw. Männer in unserem Streudiagramm), dann wendet SPSS beim Markieren folgende Logik an:

- Ist gerade *kein* Objekt markiert, bewirkt ein Mausklick auf ein beliebiges Objekt aus einer beliebigen Gruppe die Markierung aller Objekte (aus sämtlichen Gruppen).
- Ein weiterer Mausklick schränkt die Markierung auf die getroffene Gruppe ein.
- Um die Komplettmarkierung zu einer anderen Gruppe wandern zu lassen, setzt man einen Mausklick auf ein Objekt dieser Gruppe.
- Eine alternative Möglichkeit zum Markieren aller Elemente einer Gruppe ist der Mausklick auf das zugehörige Symbol in der Legende, z.B.:

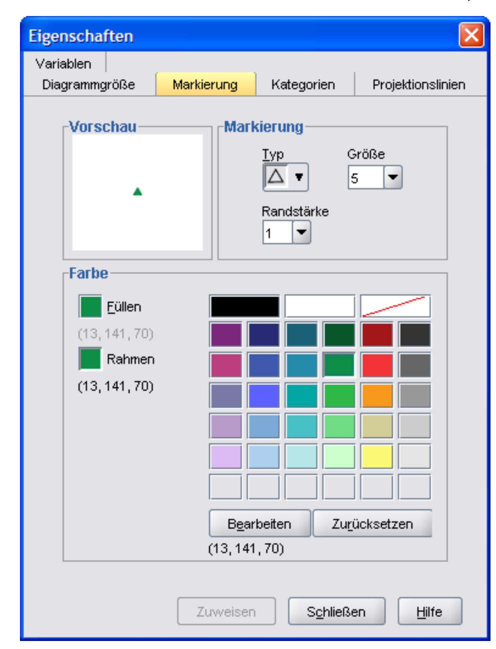


- Ist aktuell eine einzelne Gruppe markiert, kann ein einzelnes *Mitglied* dieser Gruppe durch einen Mausklick markiert werden. Alternativ gelingt die Markierung eines einzelnen Datenpunkts über das Item **Auswählen > Diese Markierung** aus seinem Kontextmenü.
- Sobald ein einzelnes Objekt markiert ist, wandert bei weiteren Mausklicks die Einzelmarkierung über Gruppengrenzen hinweg zum getroffenen Objekt.
- Bei gedrückter **Strg**-Taste ist ein gruppenunabhängiges kumulierendes Markieren möglich.
- Mit dem Lasso-Werkzeug kann man bei gedrückter linker Maustaste eine Linie um die zu markierenden Objekte (aus beliebigen Gruppen) ziehen, z.B.:





Im Beispiel liegt es nahe, für mindestens eine Gruppe nach vorangegangener Markierung ihrer Datenpunkte das zugehörige Symbol hinsichtlich Form, Größe, Randfarbe und/oder Füllfarbe zu ändern, um die beiden Gruppen besser unterscheidbar zu machen, z.B.:



Zumindest mit der deutschen SPSS-Version 17.0.3 gelingt es allerdings bei einem per **Diagrammerstellung** erzeugten gruppierten Streudiagramm oft *nicht*, den Markierungsstil für eine einzelne Gruppe zu ändern. Erstellt man dasselbe Diagramm über das veraltete Dialogfeld (wie in Abschnitt 9.1.2 beschrieben), gelingt die gruppenspezifische Änderung der Markierung problemlos.

9.2.3 Menüs und Symbolleisten

Viele Angebote sind über die Untermenüs zu den Items **Optionen** und **Elemente** im Graphikeditor-Hauptmenü sowie über äquivalente Symbolleisten verfügbar (z.B. Anpassungs- oder Interpunktionslinien, Datenbeschriftungen, Legende, Anmerkungen). Außerdem ist zu allen Objekten ein Kontextmenü verfügbar.

Im Beispiel bietet es sich an, über das Symbol oder den Menübefehl

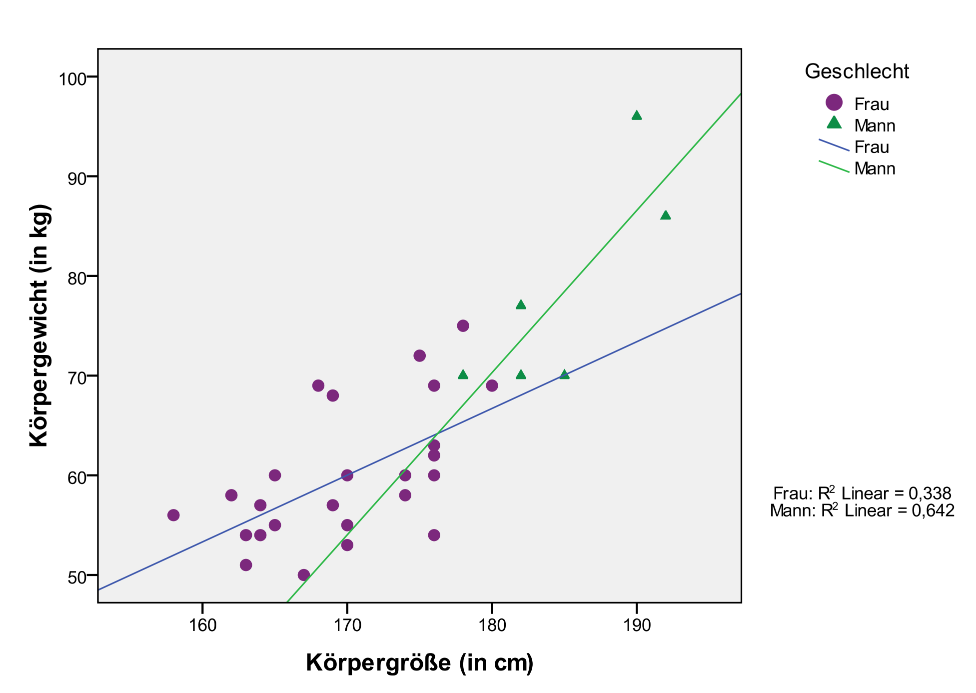
Elemente > Anpassungslinie bei Gesamtwert

die empirische Regressionsgerade einzeichnen zu lassen:

Überflüssige Objekte lassen sich über ihr Kontextmenü oder (im markierten Zustand) per **Entf**-Taste löschen. Im Beispiel könnte man so die Regressionsgerade wieder verschwinden lassen, um anschließend über das Symbol oder den Menübefehl

Elemente > Anpassungslinie bei Untergruppen

gruppenspezifische (geschlechtsbedingte) Regressionsgeraden einzufügen:



Während sich bei den Regressionsgeraden Linienstärke und –stil problemlos über das Eigenschaftsfenster ändern lassen, ist mir eine Modifikation der Farbe nicht gelungen.

Man erkennt in der Graphik einen Geschlechtsunterschied hinsichtlich der Regressionssteigung, der durch Unterschiede im Körperbau zu erklären ist:

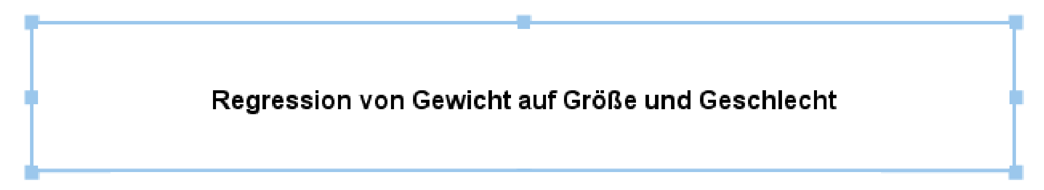


Bei zwei Männern mit 10 cm Größenunterschied ist ein stärkerer Gewichtsunterschied zu erwarten als bei zwei Frauen mit derselben Größendifferenz. Es ist also zu vermuten, dass Geschlecht den Effekt der Größe auf das Gewicht *moderiert*. Über die Analyse von Moderatoreffekten mit Hilfe der SPSS-Regressions-Prozedur informiert eine elektronische Publikation des Rechenzentrums (Baltes-Götz 2009), die auf dem Webserver der Universität Trier von der Startseite (http://www.uni-trier.de/) ausgehend folgendermaßen zu finden:

Rechenzentrum > Studierende > EDV-Dokumentationen >
Statistik > Moderatoranalyse per multipler Regression mit SPSS

9.2.4 Beschriftungen

Viele Beschriftungen (z.B. Überschriften, Legenden, Erläuterungen) besitzen nach dem Markieren einen Textrahmen mit acht Anfassern zur Größenänderung:



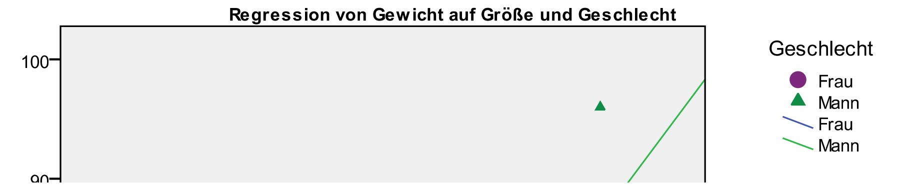
Solche Beschriftungen lassen sich auch verschieben, wobei die Transportfunktionalität des Mauszeigers am Rand aktiv wird, signalisiert durch die Zeigergestalt ✥.

Um einen Text zu ändern, markiert man ihn und setzt nach Erscheinen des Markierungsrahmens einen weiteren Mausklick darauf. Zum Beenden der Texteingabe drückt man die **Enter**-Taste oder setzt einen Mausklick außerhalb des Textrahmens.

Bei der Textformatierung kann alternativ zum Eigenschaftsfenster auch die folgende Symbolleiste verwendet werden:

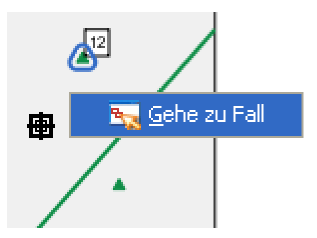


Verlässt man den Textänderungsmodus, schrumpft eventuell der Rahmen um den Text zusammen:



Um das Problem zu beheben, markiert man den Text erneut und stellt über den unteren Anfasser die gewünschte Rahmengröße her.

Über die Schaltfläche ⊞ (de)aktiviert man das Werkzeug ⊞ zur Datenbeschriftung, das zu angeklickten Datenpunkten den Wert der vereinbarten Fallbeschriftungsvariablen oder aber die laufende Datenfensterzeilennummer in die Graphik einfügt bzw. wieder entfernt, z.B.:



Nach einem rechten Mausklick auf einen Datenpunkt mit dem Fallbeschriftungswerkzeug kann man per Kontextmenü veranlassen, dass die zugehörige Zeile im Datenfenster markiert wird.

9.3 Graphiken verwenden

Wie Tabellen lassen sich auch Graphiken aus dem Ausgabefenster über die Windows-Zwischenablage in andere Anwendungen übertragen:

- Mit **Bearbeiten > Kopieren** oder **Strg+C** überträgt man eine markierte Graphik vom Ausgabefenster in die Zwischenablage.
- Mit **Bearbeiten > Einfügen** oder **Strg+V** übernimmt man sie in ein Dokument der Zielanwendung.

Als Ausgabefensterbestandteile lassen sich Graphiken sichern, drucken oder exportieren.

Zur Verwendung als **Vorlage** kann man eine Graphik aus dem Diagramm-Editor mit dem Menübefehl

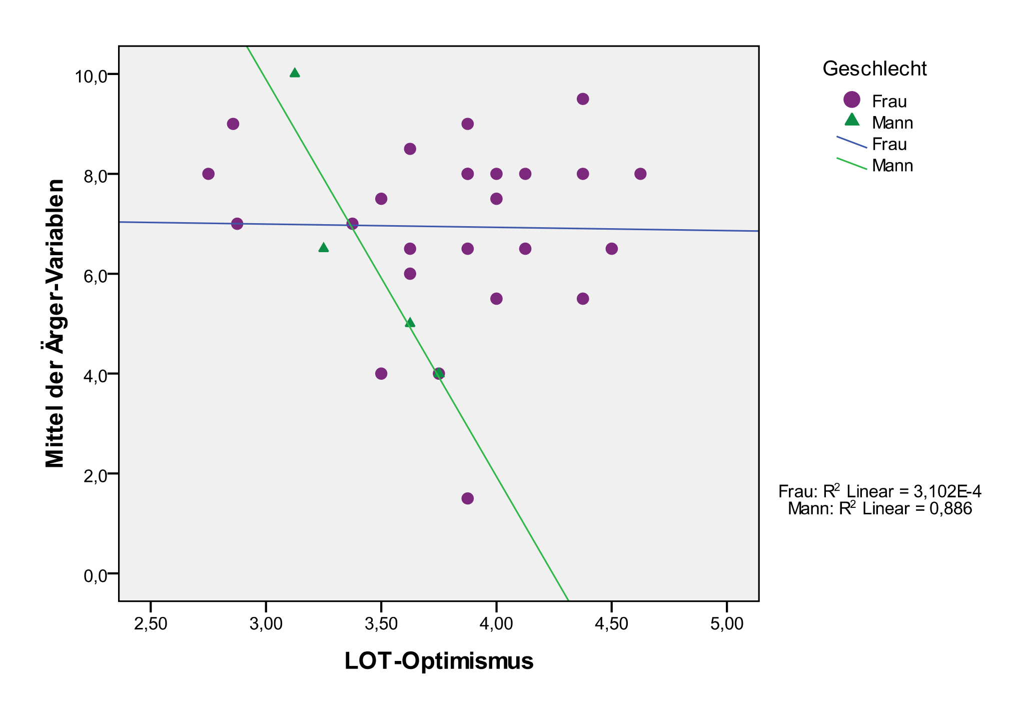
Datei > Diagrammvorlage speichern

in eine Datei mit der Namenserweiterung **sgt** sichern. Auf andere Graphiken kann man eine Vorlage bereits beim Erstellen (siehe Dialogbox **Einfaches Streudiagramm** in Abschnitt 9.1.2) oder im Diagrammeditor anwenden:

Datei > Diagrammvorlage zuweisen

9.4 Übung

Nach dem „Scheitern" der differentialpsychologischen Hypothese (siehe Abschnitt 7.4) wird man versuchen, aus den Daten Hinweise für eine Mögliche Verbesserung der Hypothese zu gewinnen. Erzeugen Sie ein Streudiagramm mit den Variablen AERGAM und LOT, und verwenden Sie wie in obigem Beispiel GESCHL als Markierungsvariable. Mit eingezeichneten Regressionsgeraden für die Untergruppen sollten Sie ungefähr folgendes Ergebnis erhalten:



Während bei den Frauen offenbar *kein* Zusammenhang zwischen LOT und AERGAM besteht, zeigt sich bei den Männern ein Effekt im Sinne unserer differentialpsychologischen Hypothese. Allerdings sollten wir die Beobachtung sehr zurückhaltend interpretieren, weil unsere Stichprobe lediglich sechs Männer enthält. Immerhin resultiert bei einer regressionsanalytischen Auswertung für den Moderatoreffekt (siehe Baltes-Götz 2009) eine relativ kleine Überschreitungswahrscheinlichkeit (0,01):

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		Regressions-koeffizient B	Standardfehler	Beta	T	Sig.
1	(Konstante)	-19,356	11,285		-1,715	,098
	Geschlecht	26,543	10,211	5,426	2,600	,015
	LOT-Optimismus	7,818	3,121	1,863	2,505	,019
	Geschlecht * LOT	-7,883	2,860	-5,633	-2,756	,010

a. Abhängige Variable: Mittel der Ärger-Variablen

Hier haben wir es aber **nicht** mit dem signifikanten Ergebnis eines statistischen Tests zu tun, sondern mit einem deskriptiven Maß zu einer interessanten Vermutung, die sich bei der explorativen Datenanalyse ergeben hat. Eine Testentscheidung über die Moderatorhypothese ist nur in einer unabhängigen Stichprobe möglich.

10 Fälle auswählen

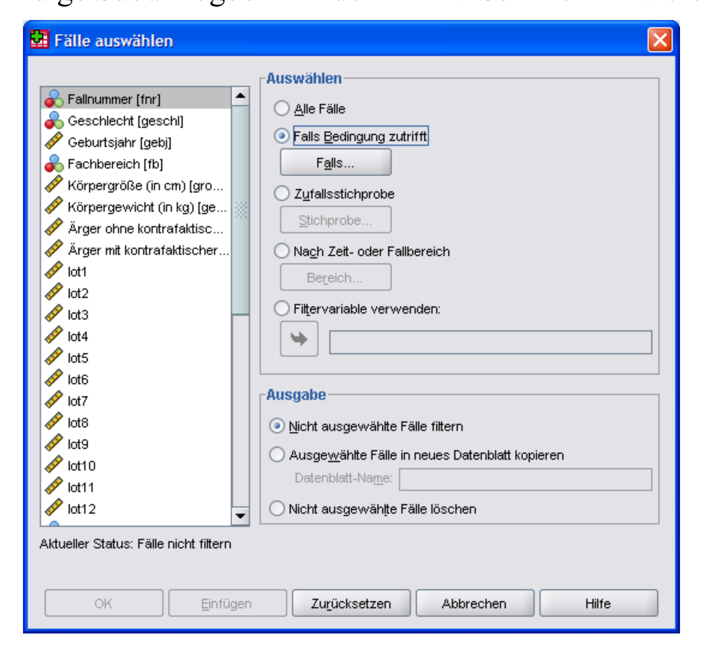
Es kommt durchaus vor, dass man sich bei einer Analyse auf eine Teilstichprobe beschränken möchte. Bei unserer KFA-Studie ist es von Interesse, die Personen mit einem *negativen* KFA-Effekt (AERGZ < 0) näher kennen zu lernen. Wir können dazu nach geeigneter Fallauswahl einen Bericht mit interessanten Variablenausprägungen anfordern.

10.1 Auswahl über eine Bedingung

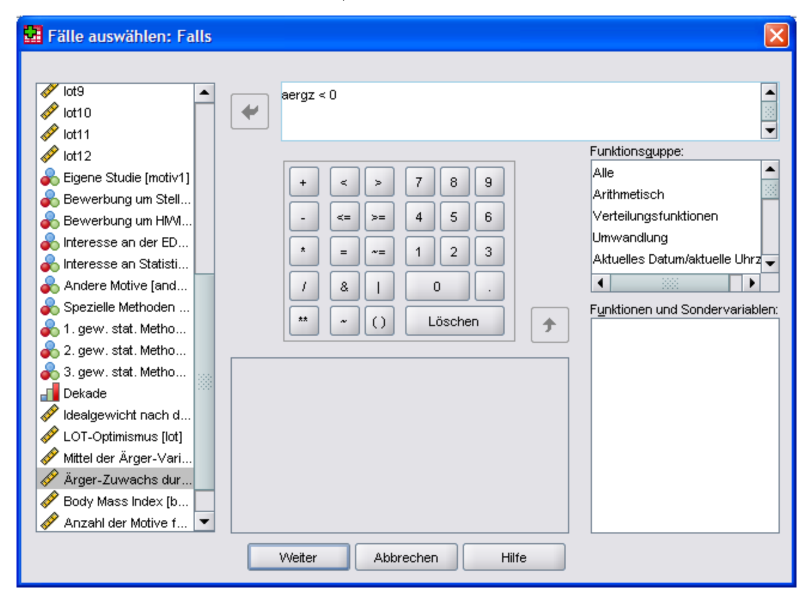
Man kann Fälle in Abhängigkeit von einer Bedingung temporär desaktivieren, aus der Arbeitsdatei entfernen oder in ein neues Datenblatt kopieren. Die zuständige Dialogbox erreichen Sie über den Menübefehl:

Daten > Fälle auswählen

Um eine Bedingung für die Teilnahme an den weiteren Auswertungen zu setzen, müssen Sie im Optionenfeld **Auswählen** die Alternative **Falls Bedingung zutrifft** markieren und anschließend die zuständige Subdialogbox mit dem **Falls**-Schalter aktivieren:



Im **Falls**-Dialogfenster haben Sie die Möglichkeit, einen logischen Ausdruck (vgl. Abschnitt 6.5.2) als Teilnahmekriterium zu definieren, z.B.:

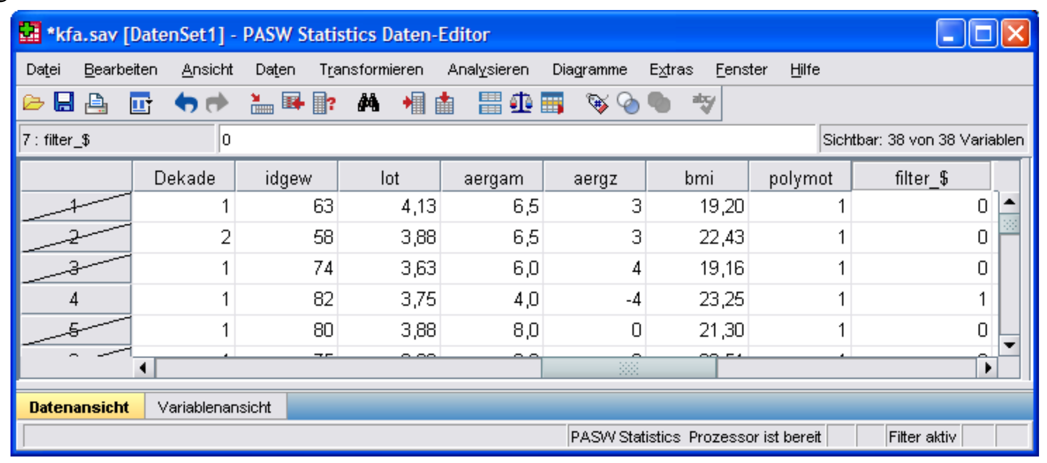


Wenn Sie nach erfolgreicher Definition des Teilnahmekriteriums **Weiter** machen, können Sie im Optionenfeld **Ausgabe** der Hauptdialogbox (siehe oben) entscheiden, was mit den Positiv- bzw. Negativ-Fällen geschehen soll:

- **Nicht ausgewählte Fälle filtern**
  SPSS erzeugt aufgrund des logischen Ausdrucks eine Hilfsvariable namens FILTER_$ mit folgenden Werten:

  1 falls bei einem Fall der logische Ausdruck wahr ist,
  0 sonst (also auch bei unbestimmtem Ausdruck).

  Diese Variable wird als **Filter** aktiviert, d.h. bis zu einer Desaktivierung des Filters werden bei allen Analysen nur noch Fälle mit dem Wert Eins bei FILTER_$ einbezogen. Die in den einstweiligen Ruhezustand versetzten Negativ-Fälle sind im Datenfenster an der durchgestrichenen Zeilennummer zu erkennen:



  Filter wirken sich nur bei statistischen und graphischen Analysen aus. Bei Datentransformationen werden auch die ausgefilterten Fälle einbezogen. Wer eine bedingte Datentransformation benötigt, muss die Methoden aus Abschnitt 6.5 verwenden.
  Wenn ein Filter aktiv ist, wird dies in der Statuszeile angezeigt (siehe Abbildung). Um den Filter später zu desaktivieren, müssen Sie die Dialogbox **Fälle auswählen** erneut aufrufen und dann im **Auswählen**-Optionenfeld den Ausgangszustand **Alle Fälle** reaktivieren.

  Per Filterkonfiguration wird die Variable FILTER_$ erstellt oder verändert. Folglich fragt SPSS am Ende der Sitzung nach, ob die veränderte Arbeitsdatei gespeichert werden soll. Wenn Sie zustimmen, landet die Variable FILTER_$ in der Datendatei. Beim nächsten Öffnen dieser Datei ist allerdings *kein* Filter aktiv. Um den durch FILTER_$ definierten Filter zu reaktivieren, muss diese Variable in der Dialogbox **Fälle auswählen** als **Filtervariable verwendet** werden. Weil Filtervariablen mit beliebigem Namen akzeptiert werden, kann man in einer SPSS-Datendatei mehrere Filtervariablen bereithalten. Außerdem kann man die einem Filter zugrunde liegende Syntax abspeichern und später wieder verwenden.

- **Ausgewählte Fälle in neues Datenblatt kopieren**
  Man erhält ein neues Datenblatt mit den Positiv-Fällen.

- **Nicht ausgewählte Fälle löschen**
  Die Negativ-Fälle werden aus der Arbeitsdatei entfernt. Aus einer eventuell zugeordneten *externen* Datei (z.B. auf der Festplatte) verschwinden die Fälle dabei *nicht*. Wenn Sie allerdings das teilentleerte Datenfenster „sichern“, haben Sie eventuell anschließend ein Problem.

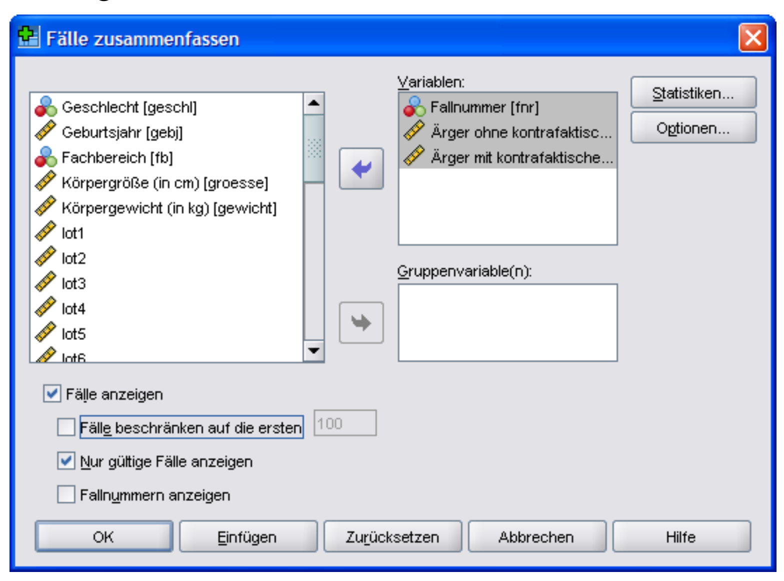
Man kann die Positiv-Fälle auch über eine **Zufallsstichprobe** gewinnen oder einen **Fallbereich** festlegen, z.B. zur Beschränkung auf die ersten *n* Fälle.

10.2 Bericht anfordern

Gelegentlich benötigt man für eine bestimmte Teilmenge von Fällen eine übersichtliche Liste mit den Ausprägungen bestimmter Variablen. Um z.B. für Personen mit negativem Ärgerzuwachs eine Liste mit den Variablen FNR, AERGO und AERGM zu erhalten, vereinbart man zunächst die Filterbedingung „AERGZ < 0" und fordert dann über

Analysieren > Berichte > Fälle zusammenfassen

die gewünschte Auflistung an:



Wir erhalten folgende Liste:

Zusammenfassung von Fällen

		Fallnummer	Ärger ohne kontrafaktische Alternative	Ärger mit kontrafaktischer Alternative
1		4	6	2
2		15	2	1
Insgesamt	N	2	2	2

11 Analyse von Kreuztabellen

Wir wollen die Hypothese prüfen, dass Frauen und Männer unterschiedliche Präferenzen bei der Wahl des Studienfachs haben.

11.1 Untersuchungsplanung

Unsere Fachbereichsvariable (FB) enthält Information über die Studienfächer der Untersuchungsteilnehmer(innen) auf einem angemessenen Aggregationsniveau. Ihre Werte stehen für die folgenden Fachbereiche der Universität Trier:

Fachbereich	Fächer
I	Pädagogik, Philosophie, Psychologie
II	Sprachorientierte Fächer
III	Historische und politische Wissenschaften
IV	BWL, Ethnologie, Informatik, Mathematik, Soziologie, VWL, Wirtsch.-Informatik
V	Jura
VI	Geowissenschaften

Nachdem die Begriffe aus der eingangs formulierten inhaltlichen Hypothese hinreichend präzisiert sind, können wir die empirisch zu prüfenden *Nullhypothese* formulieren:

Die Merkmale Geschlecht und Fachbereich sind unabhängig voneinander.

Die Unabhängigkeitsbehauptung der Nullhypothese bedeutet, dass sich aus dem Wissen über das Geschlecht eines Untersuchungsteilnehmers keinerlei Information über seine Fachbereichszugehörigkeit ableiten lässt, dass also die bedingten Fachbereichsverteilungen bei Frauen und Männern identisch sind. Zur Illustration des *Unabhängigkeits*begriffs wurde hier auf eine Verteilungs*homogenität* verwiesen. Später folgen noch einige Erläuterungen zu den beiden Begriffen und zu ihrer Beziehung.

Unsere Nullhypothesenformulierung ist „zweiseitig“, wozu es auch gar keine Alternative gibt, weil die Fachbereichsvariable mehr als zwei Stufen hat. Bei (2×2)-Kreuztabellen sind aber auch einseitige Hypothesen möglich (siehe Abschnitt 11.4.4.2).

Weil der Zusammenhang zwischen den beiden *nominalskalierten* Merkmalen Fachbereich und Geschlecht zu untersuchen ist, wählen wir als Auswertungsmethode die Kreuztabellenanalyse mit χ^2-Test. Diese Methode ist recht beliebt, wobei Einsatzfälle gelegentlich durch das wenig empfehlenswerte künstliche Kategorisieren von metrischen Variablen erzwungen werden. Hoffentlich trägt die ausführliche Behandlung der Methode im aktuellen Abschnitt nicht dazu bei, die Kreuztabellenanalyse als Universalwerkzeug der Statistik erscheinen zu lassen. Sie ist adäquat zur Untersuchung der Präferenz-Divergenz-Hypothese, weil …

- zur Aufklärung der Studienfachpräferenz nur ein einziger Prädiktor untersucht werden soll (Geschlecht),
- Kriterium und Prädiktor nominales Messniveau besitzen.

Bei sehr vielen Fragestellungen werden aber Methoden benötigt, die …

- eine beliebige Anzahl von Prädiktoren erlauben,
- das vorhandene Messniveau der Variablen unterstützen, also die enthaltene Information komplett verwerten,
- eine flexible Modellierung erlauben (z.B. Wechselwirkungen).

Mit der **logistischen Regressionsanalyse** steht für kategoriale oder ordinale Kriterien ein Verfahren bereit, das Modelle mit beliebig vielen kategorialen oder metrischen Regressoren erlaubt und auch Wechselwirkungen unterstützt (siehe z.B. Baltes-Götz 2008c).

Leider erweist sich unsere Kursstichprobe bei näherer Betrachtung als ungeeignet zur Prüfung der Präferenz-Divergenz-Hypothese, denn

- Sie ist recht klein (geringe Teststärke).
- Die Stichprobe ist wenig repräsentativ, weil nur SPSS-Interessierte enthalten sind. Folglich sind manche Fachbereiche (z.B. III, V) fast nicht vertreten.

Daher wurde eine Zufallsstichprobe der Größe n = 283 aus der Datenbank mit allen Studierenden der Universität Trier im WS 1993/94 gezogen.[1] Bei jedem Fall wurden die Variablen Geschlecht (GESCHL) und Fachbereich (FB) festgestellt. Die SPSS-Datendatei **fbgeschl.sav** mit den beiden Variablen finden Sie an der im Vorwort für Kursdateien vereinbarten Stelle.

Wir können die Stichprobengröße nicht ändern, wollen aber die daraus resultierende Power des geplanten Hypothesentests abschätzen. Dazu verwenden wir erneut das Programm **G*Power 3.1**, das schon bei der Stichprobenumfangsplanung in Abschnitt 1.3 zum Einsatz kam. Auf den Pool-PCs der Universität Trier unter dem Betriebssystem Windows ist G*Power 3.1 folgendermaßen zu starten

Start > Programme > Wissenschaftliche Programme > GPower

G*Power arbeitet bei der Kreuztabellenanalyse mit dem folgenden Effektstärkeindex W (nach Cohen 1977, S. 216)

$$W := \sqrt{\sum_{i=1}^{z}\sum_{j=1}^{s}\frac{(p_{ij}^{(1)} - p_{ij}^{(0)})^2}{p_{ij}^{(0)}}}$$

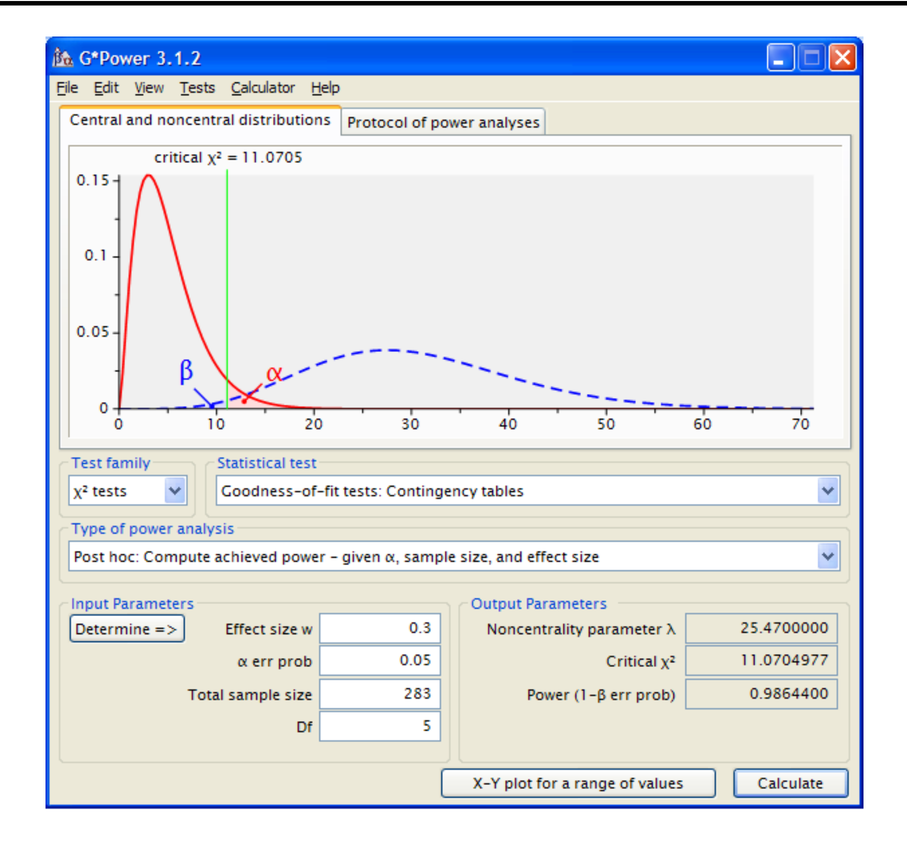
Hier werden normierte Diskrepanzen zwischen den Zellwahrscheinlichkeiten $p_{ij}^{(1)}$ unter der Alternativhypothese und den Zellwahrscheinlichkeiten $p_{ij}^{(0)}$ unter der Nullhypothese über alle Zellen aufsummiert. In Abschnitt 11.4.1 wird sich ein enger Zusammenhang zwischen dem Effektstärkeindex W und Pearsons Prüfgröße zur Unabhängigkeitshypothese sowie zu Cramers V (einem Maß der Assoziationsstärke für zwei nominalskalierte Variablen) herausstellen. Weil keine Informationen über die Effektstärke in der Population verfügbar sind, nehmen wir einen *mittleren* Wert an, per Konvention definiert durch $W = 0{,}3$.

Wir wählen in G*Power 3.1 folgende Einstellungen:

• **Test family**	**χ^2-Tests**
• **Statistical test**	**Goodness-of-fit tests Contingency tables**
• **Type of power analysis**	**Post hoc**
• **Effect size w**	0.3
• **α err prob**	0.05
• **Total sample size**	283
• **Df**	**5** Warum bei einer Tabelle mit zwei Zeilen und sechs Spalten gerade fünf Freiheitsgrade zustande kommen, erfahren Sie in Abschnitt 11.4.1.

Es resultiert eine erfreulich hohe Power von 0,99:

[1] Aufmerksame Leser(innen) werden zu Recht fragen, warum nicht *alle* Trierer Studierenden einbezogen wurden. Eine größere Stichprobe bringt stabilere Ergebnisse und hätte in dieser speziellen Situation kaum mehr „gekostet". Allerdings habe ich aus didaktischen Gründen eine Stichprobe mit „typischem" Umfang vorgezogen.

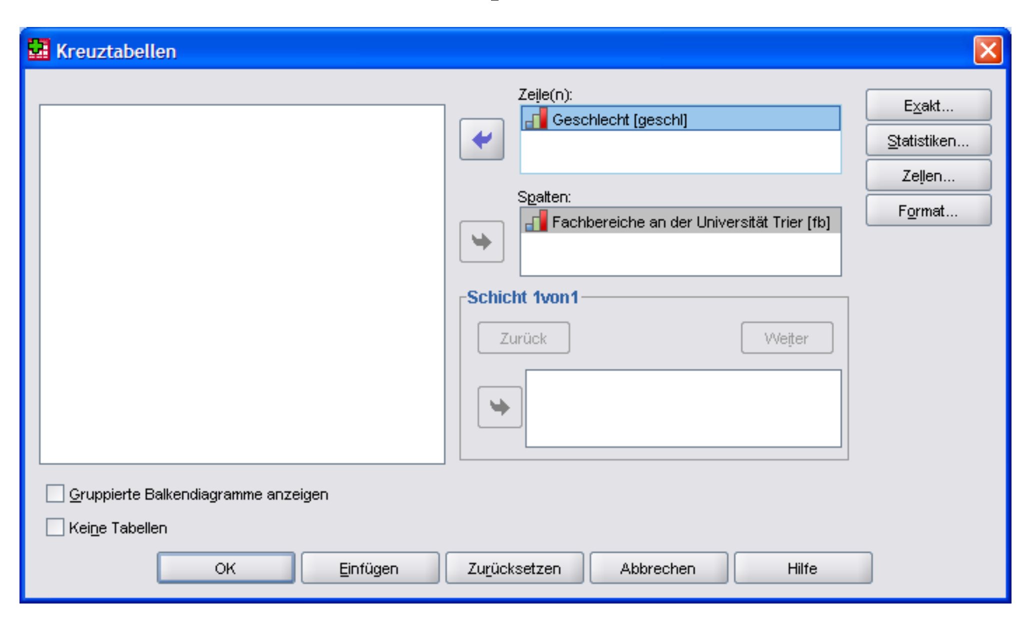


11.2 Beschreibung der bivariaten Häufigkeitsverteilung

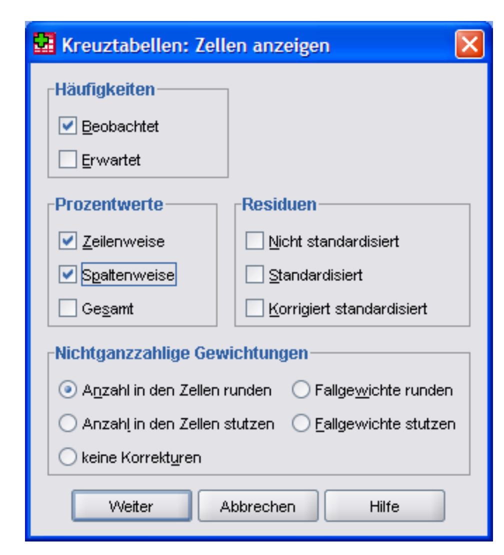
Die SPSS-Dialogbox zur Analyse zweidimensionaler Kontingenztabellen erscheint nach dem Menübefehl:

Analysieren > Deskriptive Statistiken > Kreuztabellen

Wir wählen GESCHL als Zeilen- und FB als Spaltenvariable:



In der **Zellen**-Subdialogbox kann man u.a. zeilen- und spaltenbezogene Prozentangaben für die Zellen der Kontingenztabelle anfordern:



Aufgrund dieser Spezifikationen erhalten wir für unsere Stichprobe die folgende Kreuztabelle[1]:

Geschlecht * Fachbereiche an der Universität Trier Kreuztabelle

		Fachbereiche an der Universität Trier						
		I	II	III	IV	V	VI	Gesamt
Frauen	Anzahl	29	26	18	22	26	23	144
	% von Geschlecht	20,1%	18,1%	12,5%	15,3%	18,1%	16,0%	100,0%
	% von FB	63,0%	66,7%	50,0%	31,0%	54,2%	53,5%	50,9%
Männer	Anzahl	17	13	18	49	22	20	139
	% von Geschlecht	12,2%	9,4%	12,9%	35,3%	15,8%	14,4%	100,0%
	% von FB	37,0%	33,3%	50,0%	69,0%	45,8%	46,5%	49,1%
Gesamt	Anzahl	46	39	36	71	48	43	283
	% von Geschlecht	16,3%	13,8%	12,7%	25,1%	17,0%	15,2%	100,0%
	% von FB	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Durch die Einträge in den Zellen wird die gemeinsame Verteilung der beiden Variablen GESCHL und FB beschrieben:

- Oben ... steht die absolute Häufigkeit der Zelle
  Z.B. befanden sich in der Stichprobe 17 Männer aus dem Fachbereich I.
- In der Mitte ... steht der prozentuale Anteil der Zelle an allen Fällen in der zugehörigen Zeile.
  Z.B. gehörten von den 139 *männlichen* Untersuchungsteilnehmern 12,2% zum Fachbereich I. Diese auf die Zeile bezogenen relativen Häufigkeiten beschreiben also die bedingte Verteilung der Spaltenvariablen (FB) für einen festen Wert der Zeilenvariablen (GESCHL). Wir erhalten z.B. für die Männer die folgende bedingte Verteilung der Fachbereichs-Variablen:

I	II	III	IV	V	VI
12,2%	9,4%	12,9%	35,3%	15,8%	14,4%

- Unten ... steht der prozentuale Anteil der Zelle an allen Fällen in der zugehörigen Spalte
  Z.B. waren von den 46 Personen aus dem Fachbereich I 37% Männer. Diese auf die Spalte bezogenen relativen Häufigkeiten beschreiben also die bedingte Verteilung der Zeilenvariablen (GESCHL) für einen festen Wert der Spaltenvariablen (FB). Wir erhalten z.B. für den Fachbereich I die folgende bedingte Geschlechtsverteilung:

Frauen	63%
Männer	37%

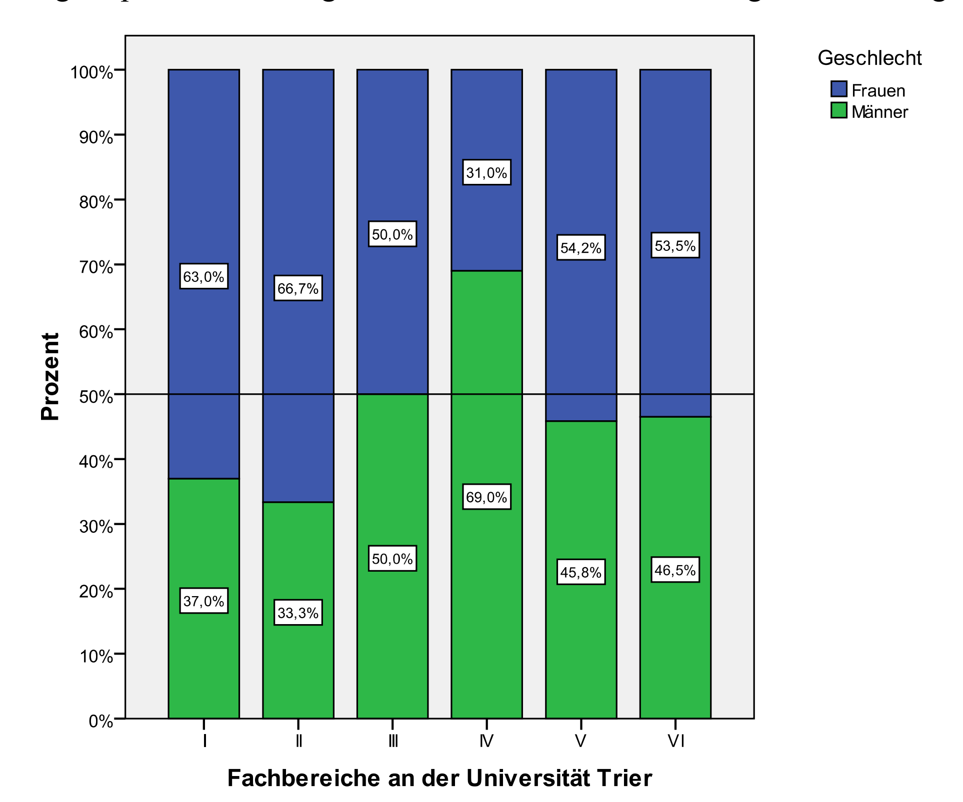
[1] Die Tabelle wurde mit dem Pivot-Editor durch Aufheben der Gruppierung Geschlecht etwas schlanker gemacht.

In der **Zellen**-Subdialogbox können auch noch weitere Informationen zu den Zellen angefordert werden (z.B. der prozentuale Anteil an der Gesamtstichprobe, die erwartete Häufigkeit unter der Nullhypothese).

Beim Vergleich der fachbereichsbedingten Geschlechtsverteilungen zeigen sich erhebliche Unterschiede:

- In den Fachbereichen I und II dominieren die Frauen mit einem Anteil von 63 bzw. 66,7%.
- Im Fachbereich IV sind die Frauen mit einem Anteil von 31% in der Minderheit.
- In den übrigen Fachbereichen III, V und VI zeigt sich ein relativ ausgeglichenes Geschlechtsverhältnis.

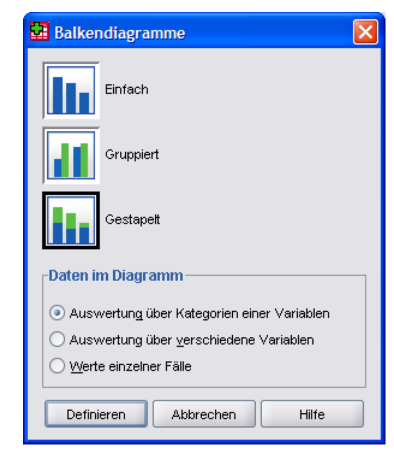
Das folgende gestapelte Balkendiagramm veranschaulicht die bedingten Verteilungen:



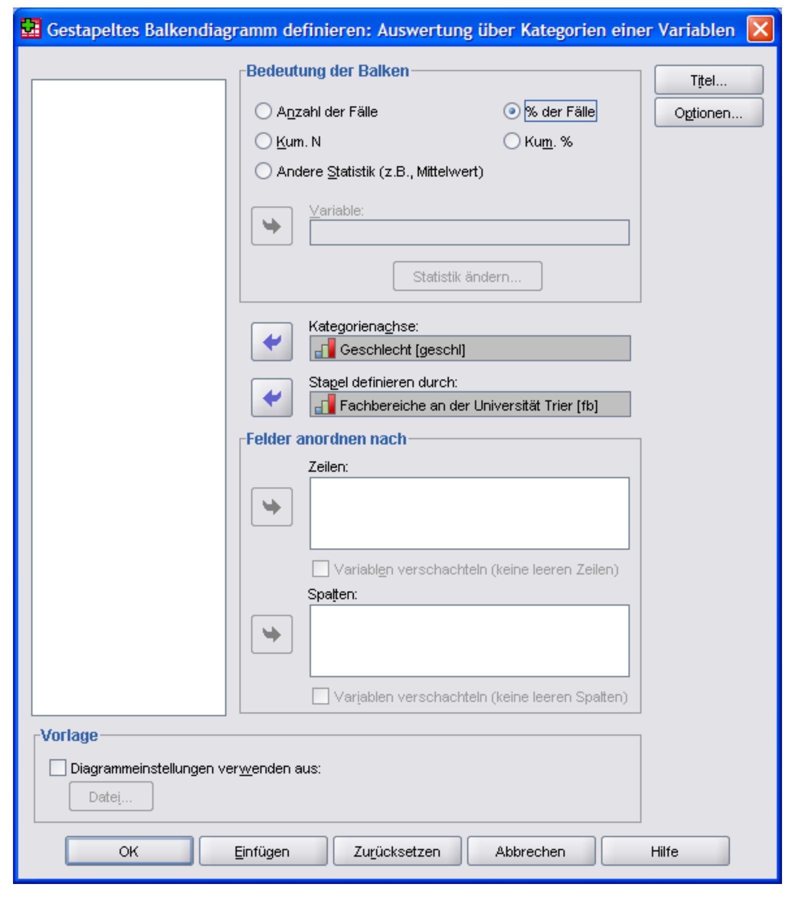
Sie können es nach dem Menübefehl

Diagramme > Veraltete Dialogfelder > Balken

und der Entscheidung für ein **gestapeltes** Balkendiagramm mit den **Kategorien einer Variablen** als **Daten im Diagramm**



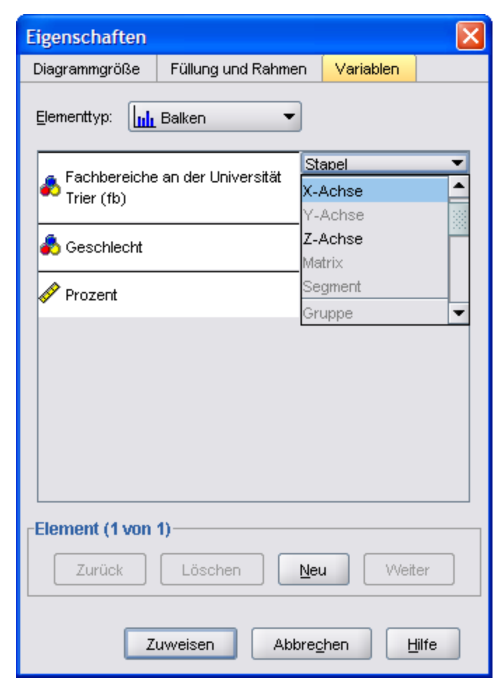
mit folgender Dialogbox anfordern:



Machen Sie **% der Fälle** zur **Bedeutung der Balken**. Indem man zunächst GESCHL als Kategorien- und FB als Stapelvariable verwendet und später die Rollen vertauscht, erzielt man den gewünschten Bezug für die Prozentangaben auf den Balken.

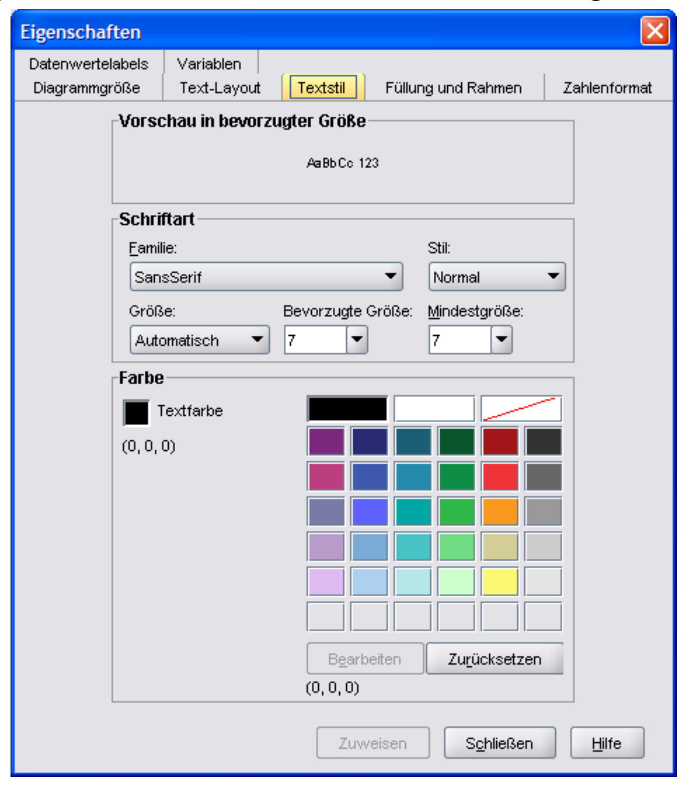
Nehmen Sie im Diagramm-Editor folgende Anpassungen vor:

- Mit Hilfe der Eigenschaftsfenster-Registerkarte **Variablen** tauschen GESCHL und FB ihre Rollen:

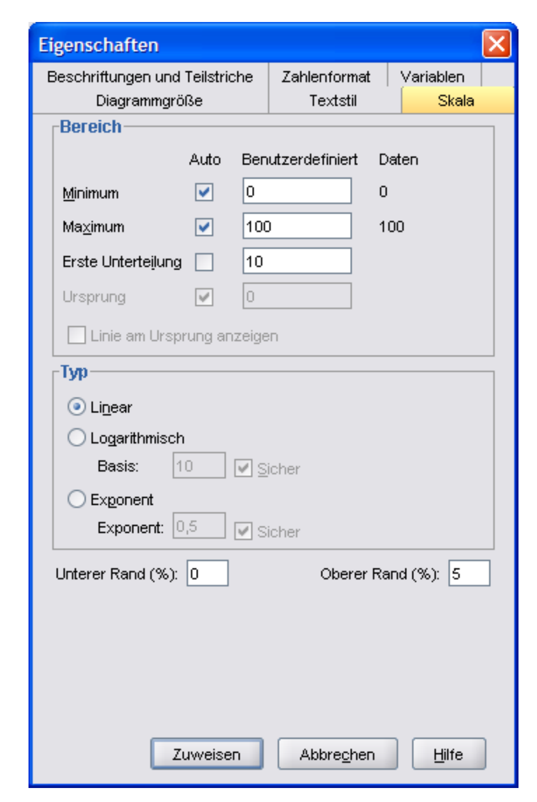


- Über **Optionen > Bezugslinie für Y-Achse** wird die 50% - Marke hervorgehoben:

- Über **Elemente > Datenbeschriftungen einblenden** oder den Symbolschalter sorgen wir bei markierten Balken für eine Anzeige der Prozentwerte. Über die Eigenschaftsfenster-Registerkarte **Textstil** erhalten diese Beschriftungen den Schriftgrad 7:



- Die Y-Achse erhält über die Eigenschaftsfenster-Registerkarte **Skala** als **erste Unterteilung** den Wert 10:



Außerdem wird auf der Registerkarte **Zahlenformat** die Dezimalstelle entfernt.

11.3 Die Unabhängigkeits- bzw. Homogenitätshypothese

Hypothesen zum Zusammenhang zwischen zwei nominalskalierten Merkmalen lassen sich auf letztlich äquivalente Weise durch Verwendung verschiedener wahrscheinlichkeitstheoretischer Begriffen formulieren. Dies soll an unserem Beispiel demonstriert werden, damit Sie die Äquivalenz verstehen und ausnutzen lernen. Es ist ja generell sinnvoll, einen Sachverhalt aus verschiedenen Blickrichtungen zu betrachten.

1. Formulierung: Unabhängigkeitshypothese

H_0: Die Merkmale Geschlecht und Fachbereich sind unabhängig, d.h. die Wahrscheinlichkeit für jedes Verbundereignis (z.B. Mann im Fachbereich V) ist gleich dem Produkt aus den Wahrscheinlichkeiten der Randereignisse (im Beispiel: Mann, Fachbereich V).

H_1: Die Merkmale Geschlecht und Fachbereich sind abhängig, d.h. die Wahrscheinlichkeit für mindestens ein Verbundereignis ist ungleich dem Produkt aus den Wahrscheinlichkeiten der Randereignisse.

2. Formulierung: Homogenitätshypothese

H_0: Die Frauenanteile sind in allen Fachbereichen gleich.

H_1: Die Frauenanteile in den Fachbereichen sind verschieden.

Man kann leicht zeigen (vgl. Hartung 1989, S. 412): Perfekte Homogenität liegt genau dann vor, wenn die Merkmale Geschlecht und Fachbereich unabhängig sind.

11.4 Testverfahren

11.4.1 Asymptotische χ^2 - Tests

Die bekannteste Prüfgröße zur Testung der Unabhängigkeits- bzw. Homogenitätshypothese ist die folgende χ^2_{P} **- Statistik nach Pearson**:

$$\chi^2_{\mathrm{P}} := \sum_{i=1}^{z} \sum_{j=1}^{s} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}, \quad \text{mit } m_{ij} := \frac{n_{i.} \cdot n_{.j}}{n}$$

Darin bedeuten:

z, s	=	Anzahl der Zeilen bzw. Spalten
n_{ij}	=	beobachtete Häufigkeit in Zelle *ij*
m_{ij}	=	geschätzte erwartete Häufigkeit in Zelle *ij* unter der H_0
$n_{i.}$	=	beobachtete Häufigkeit in Zeile *i*
$n_{.j}$	=	beobachtete Häufigkeit in Spalte *j*
n	=	Umfang der Gesamtstichprobe

Die Formel zur Schätzung der erwarteten Häufigkeiten m_{ij} unter der Nullhypothese

$$m_{ij} := \frac{n_{i.} \cdot n_{.j}}{n}$$

ist leicht nachvollziehbar. Zunächst soll die Wahrscheinlichkeit $p_{ij}^{(0)}$ der Zelle *ij* unter der H_0 bestimmt werden. Da es sich hier um ein Verbundereignis aus zwei *unabhängigen* (H_0!) Einzelereignissen handelt (Zeile *i* und Spalte *j)*, ergibt sich $p_{ij}^{(0)}$ als Produkt der Wahrscheinlichkeiten $p_{i.}$ und $p_{.j}$ für die beiden verknüpften Einzelereignisse.

$$p_{ij}^{(0)} = p_{i.} \cdot p_{.j}$$

Die Wahrscheinlichkeiten $p_{i.}$ und $p_{.j}$ sind allerdings nicht bekannt, sondern müssen durch die entsprechenden relativen Häufigkeiten in der Stichprobe geschätzt werden.[1] Die Wahrscheinlichkeit $p_{i.}$ zur Zeile *i* wird geschätzt durch die relative Häufigkeit der Zeile *i* in der Stichprobe:

$$\hat{p}_{i.} := \frac{n_{i.}}{n}$$

Analog ergibt sich die geschätzte Wahrscheinlichkeit $p_{.j}$ der Spalte *j*:

$$\hat{p}_{.j} := \frac{n_{.j}}{n}$$

Damit gilt für die geschätzte Wahrscheinlichkeit der Zelle *ij*:

$$\hat{p}_{ij}^{(0)} = \hat{p}_{i.} \cdot \hat{p}_{.j} = \frac{n_{i.}}{n}\frac{n_{.j}}{n} = \frac{n_{i.} \cdot n_{.j}}{n^2}$$

[1] Diese Formulierung geht davon aus, dass man *eine* Stichprobe gezogen und bei jedem Fall die *beiden* Merkmale Geschlecht und Fachbereich beobachtet hat. Ein anderes Stichprobenmodell läge vor, wenn man in jedem Fachbereich eine Stichprobe der festen Größe 50 gezogen und bei jedem Fall die eine Variable Geschlecht beobachtet hätte. Dann wären die Randwahrscheinlichkeiten der FB-Kategorien bekannt. Allerdings bleiben auch unter dem alternativen Stichprobenmodell alle vorgestellten Rechnungen und Entscheidungsregeln korrekt.

Die Wahrscheinlichkeit $p_{ij}^{(0)}$ lässt sich interpretieren als Erwartungswert der Indikator-Zufallsvariablen X_{ij} zur Zelle (i, j) beim Ziehen *eines* Falles:

- Tritt die Zelle (i, j) auf, nimmt X_{ij} den Wert Eins an,
- bei jedem anderen Ergebnis nimmt X_{ij} den Wert Null an.

Werden *n* Fälle unabhängig gezogen, realisieren sich *n* unabhängige Zufallsvariablen $X_{ij}^{(k)}$, $k = 1, .., n$, mit dem identischem Erwartungswert p_{ij}, und der Erwartungswert der Summenvariablen

$$\mathrm{E}(\sum_{k=1}^{n} X_{ij}^{(k)}) = \sum_{k=1}^{n} \mathrm{E}(X_{ij}^{(k)}) = n \cdot p_{ij}$$

ist die erwartete Häufigkeit der Zelle (i, j).

Mit der geschätzten Wahrscheinlichkeit $\hat{p}_{ij}^{(0)}$ ergibt sich also die geschätzte erwartete Häufigkeit m_{ij} in Pearsons Teststatistik:

$$m_{ij} = n \cdot \hat{p}_{ij}^{(0)} = n \cdot \frac{n_{i.} \cdot n_{.j}}{n^2} = \frac{n_{i.} \cdot n_{.j}}{n}$$

In Pearsons χ^2_{P}-Statistik werden die quadrierten Abweichungen der beobachteten Häufigkeiten von den geschätzten Erwartungswerten unter der H_0 aufsummiert. Durch das Quadrieren werden größere Diskrepanzen besonders stark gewichtet. Jede quadrierte Abweichung wird außerdem *normiert*, indem sie durch ihren erwarteten Wert dividiert wird. Steht etwa dem erwarteten Wert 5 die beobachtete Häufigkeit 15 gegenüber, so resultiert die quadrierte und normierte Diskrepanz 20:

$$\frac{(15-5)^2}{5} = 20$$

Dieselbe Abweichung einer beobachteten Häufigkeit 2010 vom erwarteten Wert 2000 erbringt jedoch sinnvollerweise nur eine quadrierte und normierte Diskrepanz von 0,05:

$$\frac{(2010-2000)^2}{2000} = 0{,}05$$

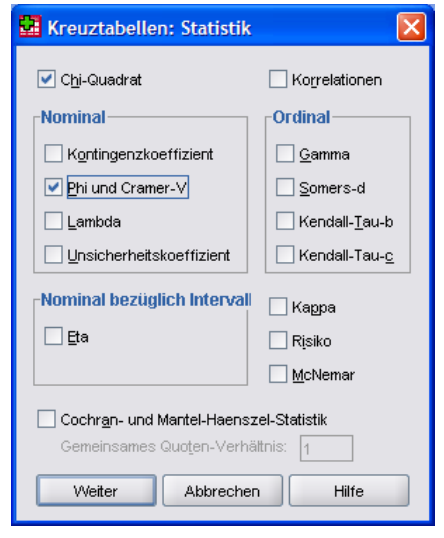
Der χ^2_{P}-Wert ist offenbar, wie es in Abschnitt 7.1 von einer Teststatistik gefordert wird, indikativ für Abweichungen von der Nullhypothese. Mit $\frac{n_{ij}}{n}$ als geschätzter Wahrscheinlichkeit $\hat{p}_{ij}^{(1)}$ der Zelle (i, j) unter der Alternativhypothese (beliebige Multinomialverteilung der Häufigkeiten in den $z{\cdot}s$ Zellen) und $\frac{m_{ij}}{n}$ als geschätzter Wahrscheinlichkeit $\hat{p}_{ij}^{(0)}$ der Zelle (i, j) unter der Nullhypothese zeigt sich sogar ein sehr enger Bezug zwischen Pearsons χ^2_{P}-Prüfgröße und dem Effektstärkeindex *W* (vgl. Abschnitt 11.1):

$$\chi^2_{\mathrm{P}} = \sum_{i=1}^{z}\sum_{j=1}^{s} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} = n\sum_{i=1}^{z}\sum_{j=1}^{s} \frac{(\frac{n_{ij}}{n} - \frac{m_{ij}}{n})^2}{\frac{m_{ij}}{n}} = n\sum_{i=1}^{z}\sum_{j=1}^{s} \frac{(\hat{p}_{ij}^{(1)} - \hat{p}_{ij}^{(0)})^2}{\hat{p}_{ij}^{(0)}} = n\hat{W}^2$$

Außerdem erfüllt die χ_P^2-Teststatistik nach Pearson auch die Verteilungsbedingung aus Abschnitt 7.1, wenn auch nur approximativ. Unter der Nullhypothese ist die χ_P^2-Statistik asymptotisch, d.h. für $n \to \infty$, χ^2-verteilt mit $df = (z - 1) \cdot (s - 1)$ Freiheitsgraden.[1] Für unsere Kreuztabelle erhalten wir also: $df = 1 \cdot 5 = 5$.

Folglich kann mit Pearsons χ_P^2-Statistik nicht nur die Plausibilität der H_0 deskriptiv beurteilt werden, sondern es kann eine empirische Überschreitungswahrscheinlichkeit berechnet und nach den Regeln aus Abschnitt 7.1 ein Signifikanztest durchgeführt werden.

In SPSS wird die χ_P^2-Statistik samt Signifikanztest mit dem Kontrollkästchen **Chi-Quadrat** in der **Kreuztabellen**-Subdialogbox **Statistik** angefordert:



Zur Beurteilung der empirischen Effektstärke wählen wir zusätzlich **Phi und Cramer-V** (siehe Abschnitt 11.4.2).

Wir erhalten folgende Testergebnisse:

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	18,191[a]	5	,003
Likelihood-Quotient	18,570	5	,002
Zusammenhang linear-mit-linear	3,197	1	,074
Anzahl der gültigen Fälle	283		

a. 0 Zellen (,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 17,68.

Es ergibt sich ein χ_P^2-Wert von ca. 18,19, der bei $df = 5$ unter der H_0 eine Überschreitungswahrscheinlichkeit (**Asymptotische Signifikanz**) von ca. 0,003 hat, d.h. ein χ_P^2 - Wert $\geq$ 18,19 bei $df = 5$ ist unter der H_0 wenig wahrscheinlich. Insbesondere ist die empirisch ermittelte Überschreitungswahrscheinlichkeit deutlich kleiner als die üblicherweise akzeptierte Irrtums-

[1] In diesem Satz treten zwei Symbole mit ähnlicher Gestalt aber deutlich verschiedener Bedeutung auf: χ_P^2 steht für eine (letztlich heuristisch definierte) Prüfgröße, mit χ^2 ist hingegen eine theoretische Verteilung gemeint.

wahrscheinlichkeit von α = 0,05. Folglich entscheidet sich der χ^2_P - Test klar gegen die H_0. In Abschnitt 7.1 wurde dieses Argumentationsmuster der Inferenzstatistik ausführlich erläutert.

Neben der χ^2_P-Statistik nach Pearson berechnet SPSS noch die alternative Prüfgröße χ^2_{LQ}, die auf dem **Likelihood-Quotienten - Prinzip** basiert. Letztere ist unter der H_0 ebenfalls asymptotisch, d.h. für $n \to \infty$, χ^2 - verteilt mit $df = (z\text{-}1)\cdot(s\text{-}1)$ Freiheitsgraden, und trotz unterschiedlicher Herleitung sind beide Statistiken asymptotisch äquivalent, d.h. mit wachsender Stichprobengröße werden sie immer ähnlicher. Während bei größeren Stichproben wegen der asymptotischen Äquivalenz die Entscheidung für eine der beiden Prüfgrößen beliebig ist, sprechen einige Befunde dafür, bei kleineren Stichproben die χ^2_P-Statistik nach Pearson wegen der besseren Verteilungsapproximation zu bevorzugen (siehe z.B. Hartung 1989, S. 439). Damit ist es also vertretbar, die χ^2_P-Statistik nach Pearson grundsätzlich gegenüber der Likelihood-Quotienten - Prüfgröße zu bevorzugen. SPSS liefert stets beide Prüfgrößen. In unserem Fall sind die Unterschiede geringfügig und für die Testentscheidung irrelevant.

Die Pearson- und die Likelihood-Quotienten-Statistik zur Beurteilung der Unabhängigkeits- bzw. Homogenitätshypothese sind nur **asymptotisch**, d.h. für $n \to \infty$, χ^2-verteilt. Für die Zulässigkeit der zugehörigen Hypothesentests setzt man üblicherweise voraus, dass alle **erwarteten** Häufigkeiten m_{ij} mindestens gleich 5 sind. SPSS protokolliert daher für jede Kreuztabelle die minimale erwartete Häufigkeit. In unserem Fall beträgt sie 17,682, so dass keine Einwände gegen Tests auf Basis der χ^2_P- bzw. χ^2_{LQ}-Statistik bestehen.

Manche Autoren formulieren etwas abgeschwächte Voraussetzungen für die erwarteten Häufigkeiten. Siegel (1976, S. 107) verlangt z.B. für χ^2_P-Tests mit $df > 1$, dass die beiden folgenden Bedingungen erfüllt sind:

- Weniger als 20% der Zellen haben eine erwartete Häufigkeit kleiner als Fünf.
- Keine Zelle hat eine erwartete Häufigkeit kleiner als Eins.

Neben den beiden Statistiken zur Prüfung der Unabhängigkeits- bzw. Homogenitätshypothese liefert SPSS unter der Bezeichnung **Zusammenhang linear-mit-linear** auch noch den χ^2_{MH}-Wert nach **Mantel-Haenszel** zur Beurteilung der **linearen** Beziehung zwischen den beiden Variablen. Diese Statistik darf nur dann interpretiert werden, wenn beide Variablen Intervallskalenqualität besitzen. Es handelt sich nämlich schlicht um die mit (n - 1) multiplizierte quadrierte Produkt-Moment-Korrelation zwischen den beiden Variablen:

$$\chi^2_{MH} := r^2(n-1)$$

Da wir zwei kategoriale Variablen betrachten, ist diese Statistik in unserem Fall sinnlos.

11.4.2 Schätzung der Effektstärke

Zur Beurteilung der empirischen Effektstärke wählen wir in der Kreuztabellen-Subdialogbox **Statistik** (siehe Abschnitt 11.4.1) **Phi und Cramer-V**.

Der bei (2 × 2)-Tabellen als Zusammenhangsmaß empfohlene und hier als Korrelation interpretierbare **Phi-Koeffizient** wird mit Hilfe von Pearsons χ^2_P-Statistik folgendermaßen definiert:

$$\phi := \sqrt{\frac{\chi^2_P}{n}}$$

Nach einer Rechnung aus Abschnitt 11.4.1 ist ϕ damit ein Schätzer für die Populations-Effektstärke *W* (vgl. Abschnitt 11.1):

$$\phi = \hat{W}$$

Bei einer Kreuztabelle mit

$$q := \mathrm{Min}(z, s) > 2$$

haben W und ϕ einen maximale Wert

$$\sqrt{q-1}$$

größer als Eins (Cohen 1977, Abschnitt 7.2). In der Definition von Cramers *V* wird demgegenüber für den Maximalwert Eins gesorgt:

$$V := \sqrt{\frac{\chi^2_P}{n(q-1)}} = \sqrt{\frac{\chi^2_P}{n}}\frac{1}{\sqrt{q-1}} = \hat{W}\frac{1}{\sqrt{q-1}}, \text{ mit } q := \mathrm{Min}(z,s)$$

Bei der FB-GESCHL - Analyse (mit Min(*z*, *s*) = 2) ist Cramers *V* identisch mit dem Phi-Koeffizienten, und wir erhalten für beide den Wert 0,254:

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	,254	,003
	Cramer-V	,254	,003
Anzahl der gültigen Fälle		283	

Er ist nicht weit entfernt vom Wert 0,3, den wird in Abschnitt 11.1 bei der Untersuchungsplanung für den Effektstärkeindex *W* angenommen haben.

11.4.3 Exakte Tests

Für die (2 × 2)-Kreuztabellen gibt es seit Jahrzehnten mit dem **exakten Test von Fisher** eine glänzende Alternative zu den approximativen χ^2 – Tests. Wie sein Name sagt, kommt Fishers Test ohne Approximationen aus und ist daher bei jeder Stichprobe anwendbar. Erfreulicherweise bietet SPSS exakte Tests auch für beliebige (*z* × *s*)-Kreuztabellen.

Eine ausführliche Beschreibung der statistischen Verfahren, die durch das SPSS-Zusatzmodul **Exact Tests** implementiert werden (Baltes-Götz 1998), ist auf dem Webserver der Universität Trier von der Startseite (http://www.uni-trier.de/) ausgehend folgendermaßen finden:

Rechenzentrum > Studierende > EDV-Dokumentationen >
Statistik > Exakte Tests mit SPSS

Allerdings sind die traditionellen asymptotischen Verfahren nun keinesfalls obsolet, weil der exakte Test für (*z* × *s*)-Kreuztabellen wegen seines enormen Rechenaufwandes nur für kleine Stichproben durchführbar ist. Insgesamt steht für die meisten Situationen ein angemessenes Verfahren zur Verfügung:

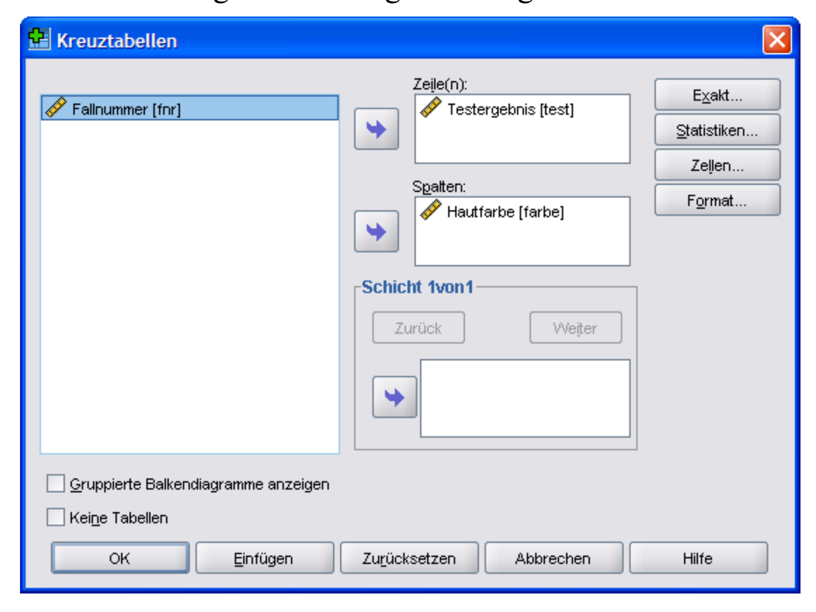
- Wenn die Anwendbarkeitskriterien für die asymptotischen Verfahren erfüllt sind, sollten Sie den Pearson-Test verwenden.
- Anderenfalls sollten Sie einen exakten Test versuchen.

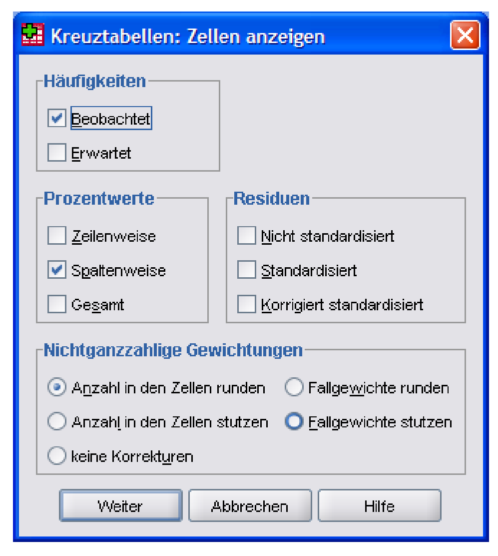
Wenn bei einer Kreuztabelle die Minimalanforderungen an die erwarteten Häufigkeiten *nicht* erfüllt sind, *und* der exakte Test aufgrund des insgesamt zu großen Stichprobenumfangs scheitert, müssen Sie die verantwortlichen schwach besetzten Zeilen bzw. Spalten entweder löschen oder miteinander bzw. mit anderen Zeilen/Spalten zusammenlegen.

In einem Anwendungsbeispiel wollen wir die Daten aus dem ersten Abschnitt des SPSS-Handbuchs zum Modul **Exact Tests** (1996, S. 1) verwenden. Es handelt sich um Prüfungsergebnisse weißer, schwarzer, asiatischer und hispanoider Feuerwehrbewerber in einer amerikanischen Kleinstadt. Diese Kreuztabelle

		Hautfarbe				
		Weiß	Schwarz	Asiatisch	Mittel- und Südamerika	Gesamt
Anzahl	Bestanden	5	2	2	0	9
	Unklar	0	1	0	1	2
	Durchgefallen	0	2	3	4	9
	Gesamt	5	5	5	5	20
Prozent	Bestanden	100,0%	40,0%	40,0%	,0%	45,0%
	Unklar	,0%	20,0%	,0%	20,0%	10,0%
	Durchgefallen	,0%	40,0%	60,0%	80,0%	45,0%
	Gesamt	100,0%	100,0%	100,0%	100,0%	100,0%

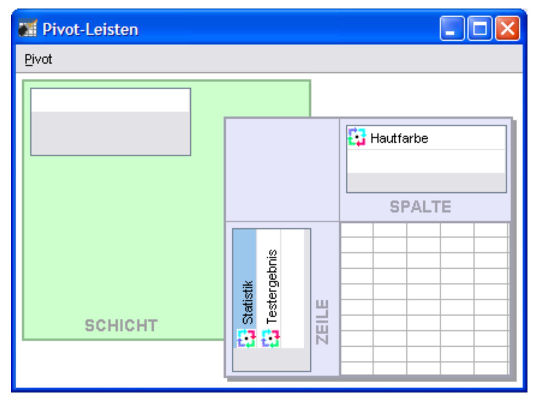
wurde mit folgenden Dialogboxen angefordert:





Der Tabellenrohling wurde per Doppelklick im Pivot-Editor geöffnet und modifiziert:

- Für die beiden Zeilendimensionen wurde per Pivot-Werkzeug die Schachtelungsordnung geändert:



- Die Gruppierungszelle zur Testergebnis-Dimension wurde durch Reduktion ihrer Breite (Verschieben des rechten Randes) zum Verschwinden gebracht.
- Die Gruppierungszelle zur Statistik-Kategorie *Prozent* hat eine neue Beschriftung erhalten.

Es soll die Nullhypothese geprüft werden, dass die Prüfungsergebnisse von der Hautfarbe unabhängig sind.

Nach einem Mausklick auf den **Exakt**-Schalter in der Dialogbox zur Kreuztabellenanalyse kann man in der folgenden Subdialogbox die **exakte** Testmethode wählen:



Daraufhin erhält man neben den approximativen Ergebnissen auch exakte Überschreitungswahrscheinlichkeiten für die Pearson- und die Likelihood-Quotienten – Teststatistik. Außerdem führt SPSS noch eine Verallgemeinerung des exakten Tests von Fisher durch, der in seiner klassischen Variante auf (2 × 2)-Tabellen beschränkt ist:

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)	Punkt-Wahrschein-lichkeit
Chi-Quadrat nach Pearson	11,556[a]	6	,073	,040		
Likelihood-Quotient	15,673	6	,016	,040		
Exakter Test nach Fisher	11,239			,040		
Zusammenhang linear-mit-linear	8,276[b]	1	,004	,004	,002	,001
Anzahl der gültigen Fälle	20					

a. 12 Zellen (100,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist ,50.

b. Die standardisierte Statistik ist 2,877.

Die approximativen χ^2 - Unabhängigkeitstests (Pearson und Likelihood-Quotient) sind nicht anwendbar, weil in allen 12 Zellen die erwartete Häufigkeit kleiner als Fünf ist. Wer dieses Problem ignoriert, andererseits aber weiß, dass der Pearson-Test gegenüber dem Likelihood-Quotienten - Test im Allgemeinen wegen der besseren Verteilungsapproximation zu bevorzugen ist, gelangt zu einer falschen Testentscheidung. Der asymptotische Pearson-χ^2 - Test empfiehlt durch eine Überschreitungswahrscheinlichkeit von 0,07, die Nullhypothese beizubehalten. Die exakte Überschreitungswahrscheinlichkeit zur Pearson-Prüfgröße beträgt hingegen 0,04, was zur Ablehnung der Nullhypothese führt.

Die exakten Überschreitungswahrscheinlichkeiten zu den drei in Frage kommenden Signifikanztests müssen nicht in jedem Fall übereinstimmen. Nachträglich die kleinste Überschreitungswahrscheinlichkeit zu wählen, ist *nicht* zulässig. Wer den Pearson-Test gemäß obiger Empfeh-

lung routinemäßig bei der allgemeinen ($z \times s$)-Kreuztabelle (mit $z \neq 2$ oder $s \neq 2$) verwendet, sofern die Zulässigkeit gegeben ist, sollte seine Prüfgröße auch bei der exakten Berechnung der Überschreitungswahrscheinlichkeit zugrunde legen.

Dass es trotz der winzigen Stichprobe zu einem signifikanten Ergebnis gereicht hat, liegt nicht nur am sensiblen Testverfahren, sondern auch an der erheblichen Effektstärke. Cramers V erreicht den Wert von 0,54:

Symmetrische Maße

		Wert	Näherungsweise Signifikanz	Exakte Signifikanz
Nominal- bzgl. Nominalmaß	Phi	,760	,073	,040
	Cramer-V	,537	,073	,040
Anzahl der gültigen Fälle		20		

Dem entspricht aufgrund des oben diskutierten Zusammenhangs

$$\hat{W} = V\sqrt{q-1}$$

eine geschätzte Effektstärke von

$$\hat{W} = 0{,}537\sqrt{3-1} = 0{,}537\sqrt{2} = 0{,}76$$

Diese ist nach Cohen (1977, S. 224f) und G*Power 3.1 als *Large* einzustufen.

11.4.4 Besonderheiten bei (2 × 2)-Tabellen

11.4.4.1 Ein klarer Fall für Fischers Test

Im beliebten Spezialfall der (2 × 2)-Tabelle ist Fishers Test nicht nur *exakt* für beliebige Stichproben, sondern er besitzt sogar unter allen „vernünftigen“, nämlich unter den so genannten unverfälschten, Tests die besten Güteeigenschaften. Daher sollten Sie in dieser Situation grundsätzlich Fishers Test verwenden. Die oben beschriebenen Rechenzeitprobleme bei exakten Tests für allgemeine ($z \times s$)-Kreuztabellen treten bei Fischers Test für die (2 × 2)-Tabelle *nicht* auf.

Für eine Teststärkeanalyse mit dem Programm G*Power 3.1 (vgl. Abschnitt 1.3.2) wählt man bei Fishers exaktem Test für die (2 × 2)-Tabelle:

- **Test family**: **Exact**
- **Statistical test**: **Proportions: … (Fisher’s exact test)**

11.4.4.2 Einseitige Hypothesen

Bei einer (2 × 2)-Tabelle lässt sich im Unterschied zu allen anderen Tabellen die Unabhängigkeits- bzw. Homogenitätshypothese auch *einseitig* formulieren. Wenn wir uns z.B. beim Vergleich der Frauenanteile unter den Studierenden der Universität Trier auf die Fachbereiche III und IV beschränken, können wir die folgende einseitige (In)homogenitätshypothese aufstellen:

H_0: Der Frauenanteil ist im FB IV mindestens genauso groß wie im FB III.

H_1: Der Frauenanteil ist im FB IV kleiner als im FB III.

Aus den (z.B. per Filterbedingung, vgl. Abschnitt 10) eingeschränkten Beispieldaten (Datei **fbgeschl.sav**) erhalten wir folgende Ergebnisse:

Kreuztabelle

	Fachbereiche an der Universität Trier		
	III	IV	Gesamt
Frauen	18	22	40
	45,0%	55,0%	100,0%
	50,0%	31,0%	37,4%
Männer	18	49	67
	26,9%	73,1%	100,0%
	50,0%	69,0%	62,6%
Gesamt	36	71	107
	33,6%	66,4%	100,0%
	100,0%	100,0%	100,0%

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	3,689[b]	1	,055		
Kontinuitätskorrektur[a]	2,922	1	,087		
Likelihood-Quotient	3,643	1	,056		
Exakter Test nach Fisher				,061	,044
Zusammenhang linear-mit-linear	3,655	1	,056		
Anzahl der gültigen Fälle	107				

a. Wird nur für eine 2x2-Tabelle berechnet

b. 0 Zellen (,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 13,46.

Wie wir bereits wissen, beträgt der Frauenanteil im FB III 50% und im FB IV 31%, die deskriptiven Statistiken fallen also klar im Sinne der Alternativhypothese aus. Der nach den obigen Überlegungen zu verwendende exakte Test von Fisher liefert für die *zweiseitige* Fragestellung eine Überschreitungswahrscheinlichkeit von 0,061, so dass die Nullhypothese beibehalten werden müsste. Bei *einseitiger* Testung erhalten wir jedoch eine Überschreitungswahrscheinlichkeit von 0,04, so dass die Nullhypothese verworfen werden kann.

Beachten Sie abschließend noch, dass sich bei Fishers Test die einseitige Überschreitungswahrscheinlichkeit keinesfalls durch Halbieren der zweiseitigen Überschreitungswahrscheinlichkeit ergibt. Die in Abschnitt 7.1 für den Spezialfall des t-Tests angegebene Regel zur Berechnung der einseitigen Überschreitungswahrscheinlichkeit aus der zweiseitigen darf also nicht generalisiert werden.

11.4.4.3 Kontinuitätskorrektur nach Yates

Bei (2 × 2)-Tabellen berechnet SPSS traditionell auch eine χ^2_Y-Größe mit Kontinuitätskorrektur nach Yates. Sie soll bei kleineren Stichproben der Pearson-χ^2_P - Statistik überlegen sein. Gemäß Abschnitt 11.4.4.1 ist sie allerdings irrelevant, weil in der (2 × 2)-Situation Fishers exakter Tests in jedem Fall vorzuziehen ist.

12 Fälle gewichten

Per Voreinstellung bezieht SPSS bei statistischen Auswertungen *alle* Fälle mit dem Gewicht *Eins* ein. In Abschnitt 10 haben Sie schon eine Möglichkeit kennen gelernt, Fälle aufgrund von Filterkriterien temporär oder permanent aus der Arbeitsdatei ausschließen. Nun erfahren Sie, wie man die Fälle individuell gewichtet, so dass sie bei statistischen Analysen unterschiedlichen Einfluss auf die Ergebnisse haben.

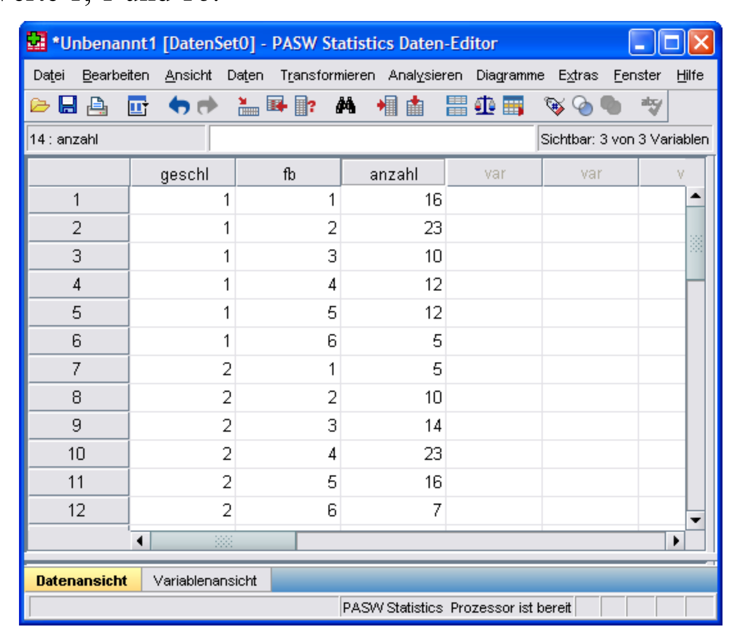
12.1 Beispiel

Die Möglichkeit, von Eins verschiedene Fallgewichte zu verwenden, d.h. z.B. einem Fall des Gewicht 16 zuzuschreiben und so zu tun, als seien 16 Fälle mit genau gleichen Variablenausprägungen in der Arbeitsdatei vorhanden, erscheint zunächst sinnlos. Aber erinnern wir uns an die (Geschlecht × Fachbereich) - Kreuztabelle aus Abschnitt 11. Zur Verwendung in einer späteren Übungsaufgabe betrachten wir hier eine strukturell identische Tabelle, die auf einer anderen Zufallsstichprobe der Größe n = 153 beruht:

	Fachbereich					
Geschlecht	I	II	III	IV	V	VI
Frau	16	23	10	12	12	5
Mann	5	10	14	23	16	7

Um mit den in Abschnitt 11 erklärten χ^2 - Tests anhand dieser Stichprobendaten prüfen zu können, ob in den Fachbereichen die Geschlechtsverteilungen verschieden sind, brauchen Sie nach unserem bisherigen Kenntnisstand eine Arbeitsdatei, in der z.B. 16 Fälle mit dem Geschlecht 1 und dem Fachbereich 1 enthalten sind, 23 Fälle mit Geschlecht 1 und Fachbereich 2 usw. Wir haben jedoch lediglich die obige Tabelle zur Verfügung. Statt nun mühselig 153 Fälle im Dateneditor neu einzutippen, können wir von der Möglichkeit der Fallgewichtung folgendermaßen Gebrauch machen:

- Wir sorgen für ein leeres Datenfenster. Dort definieren wir die Variablen GESCHL (Geschlecht), FB (Fachbereich) und ANZAHL.
- Jede Zelle der (Geschlecht × Fachbereich) - Kreuztabelle wird im SPSS-Datenfenster als *ein* Fall behandelt. Der erste Fall erhält z.B. für die drei Variablen GESCHL, FB und ANZAHL die Werte 1, 1 und 16:



	geschl	fb	anzahl	var	var
1	1	1	16		
2	1	2	23		
3	1	3	10		
4	1	4	12		
5	1	5	12		
6	1	6	5		
7	2	1	5		
8	2	2	10		
9	2	3	14		
10	2	4	23		
11	2	5	16		
12	2	6	7		

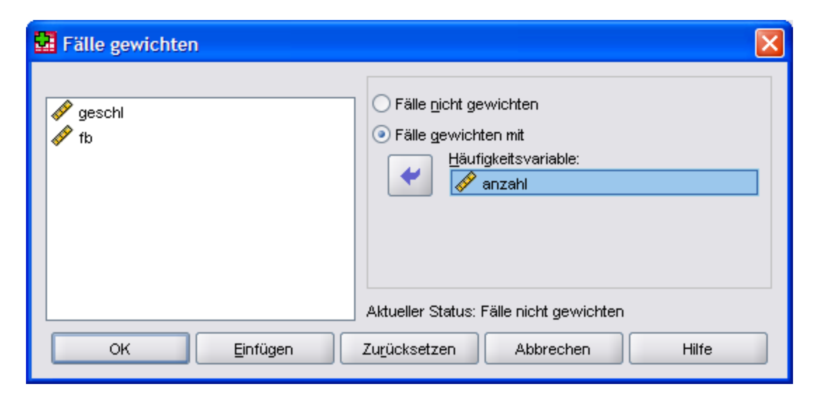
- Die Fälle werden mit der Variablen ANZAHL gewichtet. Damit tun wir z.B. so, als seien 16 Fälle mit dem Geschlecht 1 und dem Fachbereich 1 vorhanden gewesen. Aber das stimmt ja wirklich. Offenbar ist die Fallgewichtung doch nicht so sinnlos.

Um eine Gewichtsvariable zu vereinbaren, rufen wir mit dem Menübefehl

Daten > Fälle gewichten

eine Dialogbox auf, die folgende Optionen anbietet:

- **Fälle nicht gewichten**
  Damit wird eine bestehende Gewichtung wieder aufgehoben.
- **Fälle gewichten mit**
  Die gewünschte Variable wird mit dem Transportschalter in die Position der **Häufigkeitsvariablen** gebracht, z.B.:



In der Dialogbox wird außerdem angezeigt, ob momentan eine Gewichtungsvariable vereinbart ist. Dieselbe Information erscheint auch in der Statuszeile des Datenfensters (siehe oben).

Beim Einsatz von Gewichtungsvariablen ist noch zu beachten:

- Zur Gewichtung kann natürlich nur eine numerische Variable verwendet werden; diese darf allerdings auch gebrochene Werte enthalten. Negative und fehlende Werte werden auf Null gesetzt, d.h. die betroffenen Fälle werden nicht berücksichtigt, solange die Gewichtungsvariable aktiv ist.
- Ist beim Speichern der Arbeitsdatei eine Gewichtung aktiv, so wird diese mit abgespeichert und ist bei späterer Verwendung der Datendatei in Kraft.
- Bei der in diesem Abschnitt beschriebenen Anwendung der Gewichtungsoption wird dafür gesorgt, dass alle tatsächlich in der Studie vorhandenen Beobachtungen mit dem Gewicht Eins in die Kreuztabellenanalyse eingehen. Wenn die vorhandenen Beobachtungen individuelle Gewichte ($\neq 1$) erhalten, werden natürlich Signifikanztests erheblich beeinflusst. Auf jeden Fall muss dann die Gewichtungsvariable einen Mittelwert von 1 haben, d.h. die Summe der Gewichte muss gerade den Stichprobenumfang ergeben.

12.2 Übung

Prüfen Sie anhand der Daten aus der Tabelle am Anfang von Abschnitt 12.1 die Nullhypothese, dass die Merkmale Geschlecht und Fachbereich unabhängig sind.

13 Auswertung von Mehrfachwahlfragen

In Abschnitt 1.4.2.3 wurde betont, dass mit einer Mehrfachwahlfrage nicht etwa *ein* mysteriöses Merkmal mit mehreren Ausprägungen erfasst wird, wie es wohl durch manche Köpfe bzw. Alpträume spukt, sondern *eine Familie* inhaltlich verwandter dichotomer Merkmale. Eine leichte Komplikation tritt erst auf, wenn zur Vereinfachung der Erfassung ein sparsames Set aus kategorialen Variablen definiert worden ist, das für viele Auswertungen erst „ausgepackt“ werden muss.

Grundsätzlich besteht kein Bedarf für spezielle Auswertungsverfahren für die mit Mehrfachwahlfragen erfassten Variablen. Es ist allerdings gelegentlich sinnvoll, eine Häufigkeits- oder Kreuztabellenanalyse für *alle* Mitglieder einer Familie dichotomer Variablen (ob aus einer Mehrfachwahlfrage entstanden oder wie auch immer) in gleicher Form auszuführen. Für diese Situation bietet SPSS gewisse Rationalisierungsmöglichkeiten, die in diesem Abschnitt vorgestellt werden sollen. Außerdem kann SPSS für die mit einem sparsamen Set aus kategorialen Variablen erfassten dichotomen Merkmale Häufigkeits- und Kreuztabellenanalysen ohne vorheriges Auspacken durchführen.

13.1 Mehrfachantworten-Sets definieren

Im Teil 4a unseres Fragebogens haben die Teilnehmer für fünf konkrete Motive, den SPSS-Kurs zu besuchen, und eine Restkategorie alles zutreffende angekreuzt. Es liegt nahe, eine Übersicht zu erstellen, aus der für die einzelnen Motive hervorgeht, wie häufig sie gewählt worden sind. Natürlich können wir die Zustimmungsfrequenzen bei den Motiv-Variablen z.B. auch mit der längst bekannten Häufigkeitsanalyse (**Analysieren > Deskriptive Statistiken > Häufigkeiten**) bestimmen lassen. SPSS kann jedoch für solche *Gruppen zusammengehöriger Variablen* die Zustimmungshäufigkeiten sowie einige zusätzliche Ergebnisse in besonders kompakter Form ausgeben. Wir erhalten für unsere Daten die folgende Tabelle:

		N	%
Motive zur Kursteilnahme	Eigene Studie	23	76,7
	Bewerbung um Stelle	1	3,3
	Bewerbung um HIWI-Job	1	3,3
	Interesse an der EDV	5	16,7
	Interesse an Statistik	10	33,3
	Andere Motive	1	3,3

Es zeigt sich etwa, dass 23 Personen (= 76,7% von den 30 Fällen mit gültigen Werten bei den Motiv-Variablen) dem ersten Motiv zugestimmt haben. Ein Fall, auf den wir später noch eingehen müssen, fand keines der fünf konkreten Motive für sich passend und markierte die Restkategorie (*Andere Motive*).

Bei der obigen Tabelle wird die **Variablengruppe** $MOTIVE verwendet, die zuvor definiert werden muss. Wählen Sie dazu den Menübefehl:

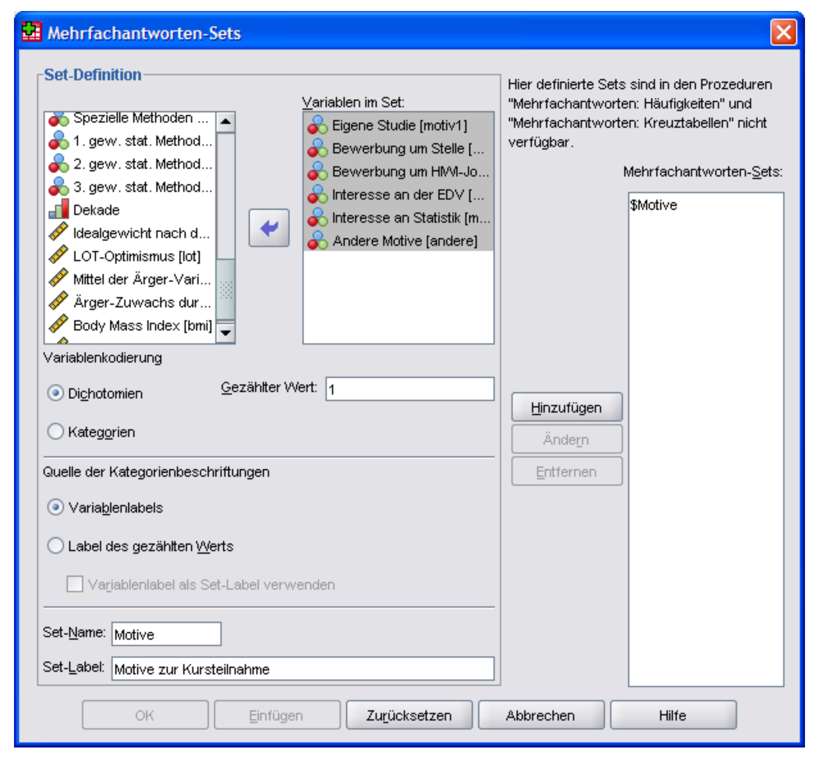
Analysieren > Tabellen > Mehrfachantworten-Sets

In der nun erscheinenden Dialogbox sind folgende Aktionen nötig:

- Befördern Sie die Variablen MOTIV1 bis MOTIV5 sowie ANDERE in die Liste **Variablen im Set**.
- Wählen Sie im Rahmen **Variablenkodierung** die Option **Dichotomien** mit dem **Gezählten Wert** Eins.

- Vereinbaren Sie für das Set den **Namen** *Motive* und das **Label** *Motive zur Kursteilnahme*.

Danach müsste Ihre Dialogbox ungefähr so aussehen:



Nehmen Sie mit **Hinzufügen** das neue Set in die Liste der **Mehrfachantworten-Sets** auf, und quittieren Sie die Dialogbox mit **OK**.

Auf die beschriebene Weise definierte Mehrfachantworten-Sets werden in der Arbeitsdatei gespeichert und ggf. in die zugeordnete Datendatei gesichert, so dass sie beim späteren Öffnen der Datei wieder zur Verfügung stehen.

Bei der Set-Definition kommt das SPSS-Kommando MRSETS zum Einsatz, das mit Hilfe der Dialogbox **Mehrfachantworten-Sets definieren** über den Schalter **Einfügen** erzeugt werden kann, z.B.:

```
MRSETS
  /MDGROUP NAME=$Motive LABEL='Motive zur Kursteilnahme'
   CATEGORYLABELS=VARLABELS
   VARIABLES=motiv1 motiv2 motiv3 motiv4 motiv5 andere VALUE=1
  /DISPLAY NAME=[$Motive].
```

Bei wichtigen Sets sollte das definierende MRSETS-Kommando in das Transformationsprogramm zum Erstellen der Fertigdatendatei aufgenommen werden (vgl. Abschnitte 6.1.1 und 6.7).

Über den Menübefehl

Analysieren > Mehrfachantworten > Variablen-Sets definieren

bzw. den zugehörigen Befehl MULT RESPONSE

```
MULT RESPONSE
  GROUPS=$Motive 'Motive zur Kursteilnahme' (motiv1 motiv2 motiv3 motiv4
  motiv5 andere (1))
  /FREQUENCIES=$Motive  .
```

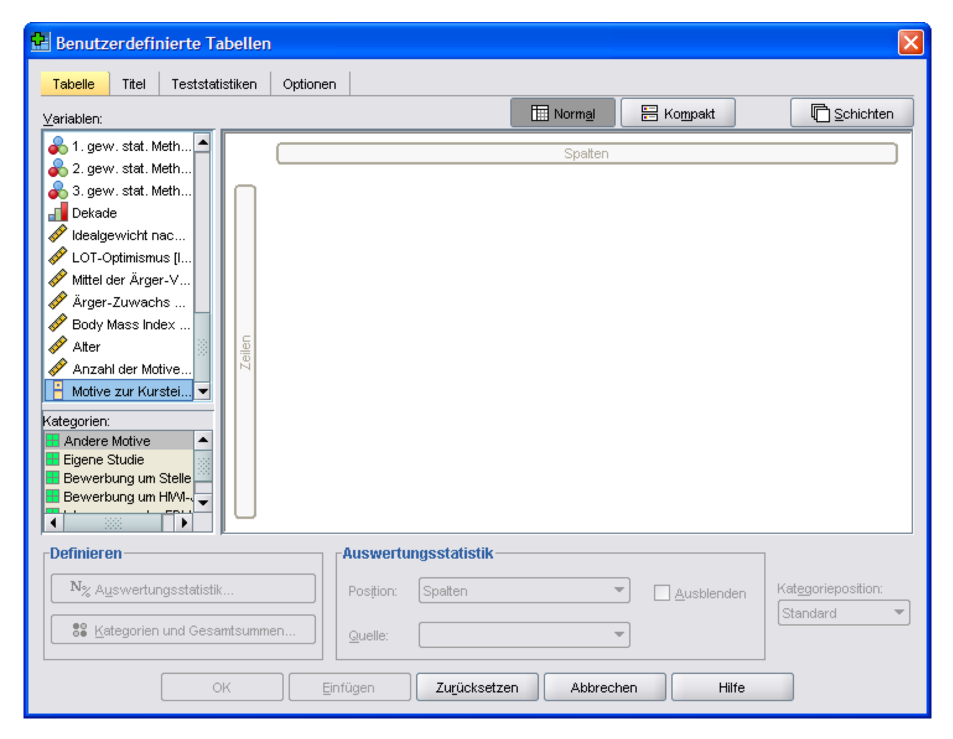
ist noch eine ältere Möglichkeit zur Set-Definition verfügbar. Ihr entscheidender Nachteil im Vergleich zur oben beschriebenen Lösung besteht darin, dass die Set-Definitionen beim Schließen des zugehörigen Datensets verschwinden, also *nicht* in einer Datendatei gespeichert werden können.

13.2 Häufigkeitstabellen für Mehrfachantworten-Sets

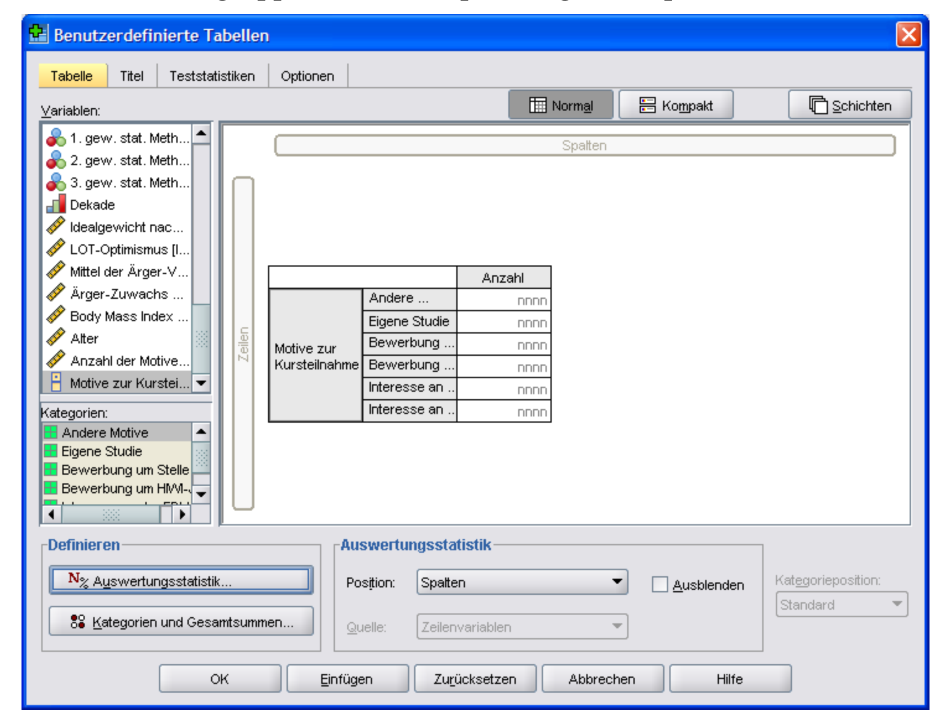
Unter Verwendung des Variablengruppe $MOTIVE (erzeugt per MRSETS-Kommando) lässt sich die in Abschnitt 13.1 präsentierte Tabelle mit den Häufigkeitsverteilungen der Set-Variablen über den Menübefehl

Analysieren > Tabellen > Benutzerdefinierte Tabellen

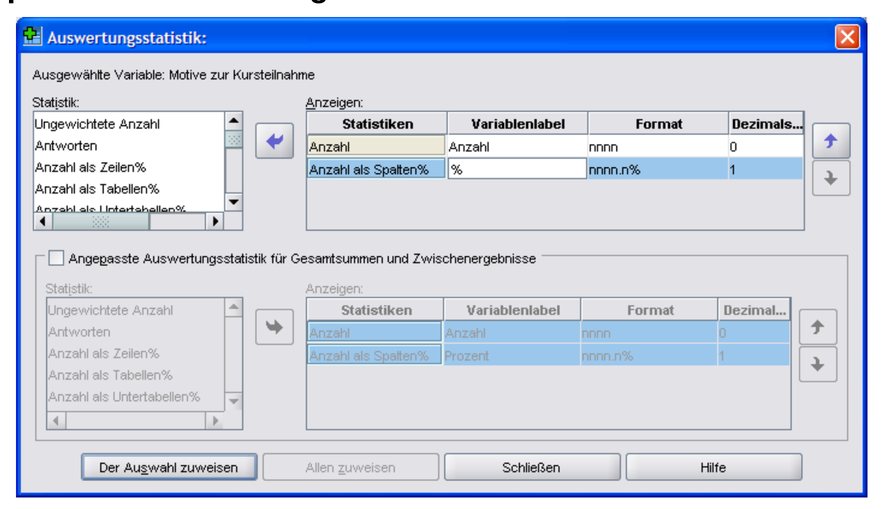
und den folgenden, analog zur Diagrammerstellung (vgl. Abschnitt 9.1.1) konstruierten Dialog anfordern:



Wir befördern die Variablengruppe $MOTIVE per Drag & Drop auf die **Zeilen**-Ablagezone:



Um die voreingestellte Häufigkeitsspalte durch eine Prozentspalte zu ergänzen, klicken wir auf den Schalter **Auswertungsstatistik** und befördern aus der **Statistik**-Liste das Element **Anzahl als Spalten%** in den **Anzeigen**-Bereich:



Wie ändern das vorgeschlagene **Variablenlabel** und quittieren mit dem Schalter **Der Auswahl zuweisen**. Wird auch noch der Dialog **Benutzerdefinierte Tabellen** mit **OK** quittiert, erscheint die fertige Tabelle im Ausgabefenster.

Entfernt man die Variable ANDERE zur Restkategorie der sonstigen Motive aus dem Set $MOTIVE, dann resultieren folgende Ergebnisse mit abweichenden Prozentwerten:

		N	%
Motive zur Kursteilnahme	Eigene Studie	23	79,3
	Bewerbung um Stelle	1	3,4
	Bewerbung um HIWI-Job	1	3,4
	Interesse an der EDV	5	17,2
	Interesse an Statistik	10	34,5

Des Rätsels Lösung ist eine SPSS-Eigenart bei der Analyse von Mehrfachwahl-Sets aus dichotomen Variablen: Als gültig werden nur solche Fälle betrachtet, die bei mindestens einer Set-Variablen den zu zählenden Wert besitzen (bei uns also die Eins). Daher wird neben dem Fall 13 mit SYSMIS bei den Variablen MOTIV1 bis MOTIV5 auch der dritte Fall ausgeschlossen, der *alle konkreten Motive verneint*, aber die Restkategorie markiert hat. Wenn SPSS in obiger Ausgabe z.B. zum Motiv Eins meldet, dass 79,3% der Fälle (23 von 29) zugestimmt hätten, ist dies schlicht falsch.

13.3 Kreuztabellen für Mehrfachantworten-Sets

Wenn wir uns für Geschlechtsunterschiede bei der Zustimmung zu den fünf konkreten Motiven interessieren (z.B.: *Wer interessiert sich mehr für Statistik?*), sind genau *fünf* (2×2)-Tabellen zu analysieren. Über den aus Abschnitt 11 bekannten Menübefehl **Analysieren > Deskriptive Statistiken > Kreuztabellen** erhalten wir z.B. für das Statistik-Motiv folgendes Ergebnis:

Interesse an Statistik * Geschlecht Kreuztabelle

			Geschlecht		Gesamt
			Frau	Mann	
Interesse an Statistik	Nein	Anzahl	15	5	20
		% von Interesse an Statistik	75,0%	25,0%	100,0%
		% von Geschlecht	62,5%	83,3%	66,7%
	Ja	Anzahl	9	1	10
		% von Interesse an Statistik	90,0%	10,0%	100,0%
		% von Geschlecht	37,5%	16,7%	33,3%
Gesamt		Anzahl	24	6	30
		% von Interesse an Statistik	80,0%	20,0%	100,0%
		% von Geschlecht	100,0%	100,0%	100,0%

Weil die Motiv-Variablen nur zwei Ausprägungen haben, sind die Ergebnisse zur Nein-Kategorie überflüssig. Es genügt zu wissen, dass 37,5% von den 24 Frauen und 16,7% von den sechs Männern ein Interesse an Statistik angegeben haben. Durch Verzicht auf die redundanten Zeilen erhält man eine sehr kompakte Darstellung der (2×2)-Tabellen zu Geschlechtsunterschieden bei den Kursmotiven:

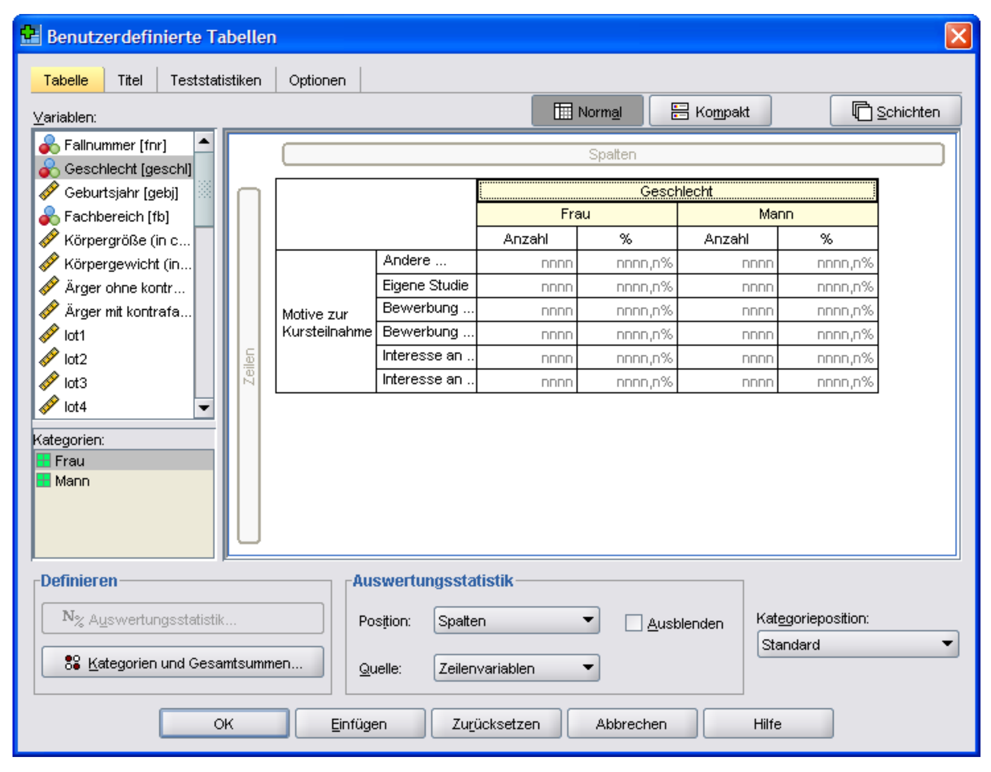
		Geschlecht					
		Frau		Mann		Gesamt	
		Anzahl	%	Anzahl	%	Anzahl	%
Motive zur Kursteilnahme	Eigene Studie	19	79,2%	4	66,7%	23	76,7%
	Bewerbung um Stelle	1	4,2%	0	,0%	1	3,3%
	Bewerbung um HIWI-Job	0	,0%	1	16,7%	1	3,3%
	Interesse an der EDV	3	12,5%	2	33,3%	5	16,7%
	Interesse an Statistik	9	37,5%	1	16,7%	10	33,3%
	Andere Motive	1	4,2%	0	,0%	1	3,3%
	Gesamt	24	100,0%	6	100,0%	30	100,0%

Beachten Sie bitte: Dies ist **nicht eine** (6×2)-Kontingenztabelle, **sondern dies sind sechs** (2×2)-Kontingenztabellen. In der vorletzten Zeile befindet sich etwa die Essenz der MOTIV5 × GESCHL - Kontingenztabelle.

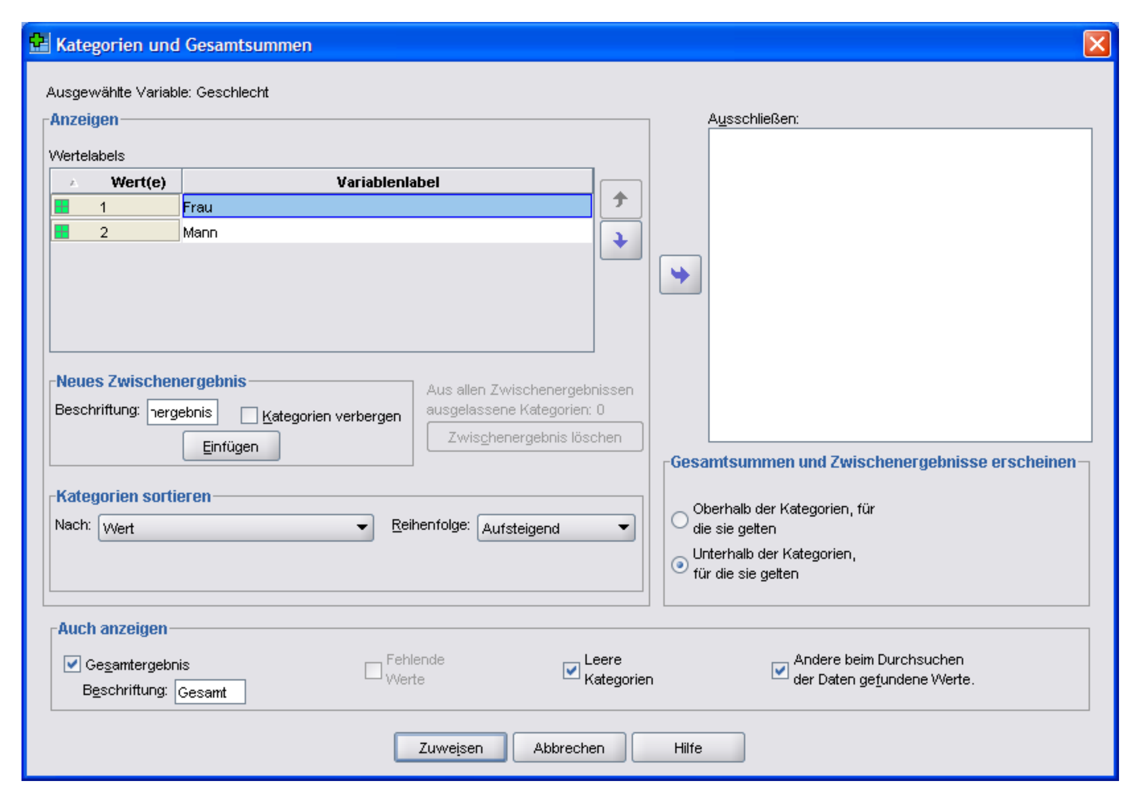
Um die obige Tabelle anzufordern, öffnen wir über

Analysieren > Tabellen > Benutzerdefinierte Tabellen

erneut die Dialogbox für benutzerdefinierte Tabellen. Ausgehend von dem in Abschnitt 13.2 erreichten Bearbeitungszustand bewegen wir die Variable GESCHL auf die **Spalten**-Ablagezone:

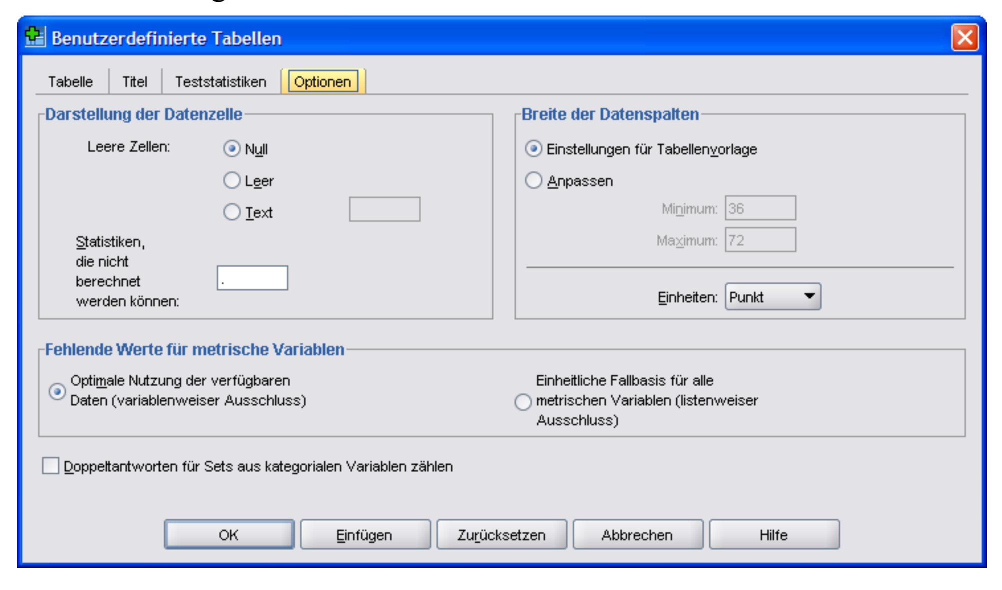


Bei markierter Zeilen- bzw. Spaltendimension öffnen wir jeweils über den Schalter **Kategorien und Gesamtsummen** den folgenden Dialog



und markieren das Kontrollkästchen **Gesamtergebnis**.

Auf der Registerkarte **Optionen** entscheiden wir uns dafür, auch die Häufigkeit **Null** explizit in betroffene Zellen einzutragen:



Auch bei den Kreuztabellen ist die in Abschnitt 13.2 kritisierte MD-Konzeption der SPSS-Mehrfachwahl-Auswertung zu beachten. Wäre nicht die Variable ANDERE Mitglied im Set $MOTIVE, dann würde SPSS in der Kombitabelle nur noch diejenigen Fälle berücksichtigen, die mindestens ein konkret abgefragtes Motiv bejaht haben.

13.4 Ein sparsames Set kategorialer Variablen expandieren

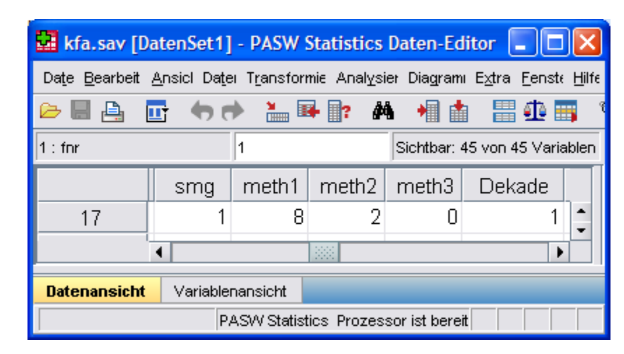
In Abschnitt 1.4.2.3 wurde das sparsame Set aus kategorialen Variablen für Mehrfachwahlfragen mit sehr vielen Antwortmöglichkeiten zur Vereinfachung der Erfassung empfohlen. Zwar ist diese Datenstruktur kein Nachteil bei den Analyseprozeduren, die in den Abschnitten 13.2 und 13.3 beschrieben wurden, doch sind Auswertungen denkbar, die ein vollständiges Set aus

dichotomen Variablen erfordern. In dieser Situation kann man das sparsame Set mit Hilfe der SPSS-Kommandosprache „expandieren“. Die folgenden Kommandos erzeugen zu unseren Variablen METH1 bis METH3 die acht dichotomen Variablen STAT1 bis STAT8, die für jeweils eine bestimmte statistische Methode festhalten, ob sie genannt worden ist (Wert Eins) oder nicht (Wert Null):

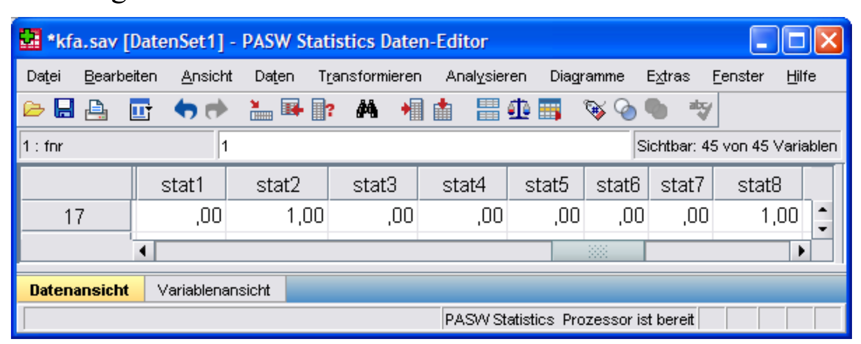
```
do repeat  stat = stat1 to stat8 /n = 1 to 8.
 do if (meth1 = n) or (meth2 = n) or (meth3 = n).
  compute stat = 1.
 else.
  compute stat = 0.
 end if.
end repeat.
execute.
```

Die Variable STAT2 steht z.B. für die Regressionsanalyse, weil gemäß Kodierplan bei einer der Variablen METH1 bis METH3 eine 2 zu notieren war, wenn ein Fall im Fragebogenteil 4b die Regressionsanalyse genannt hatte.

Beim Fall Nr. 17 wurden die genannten Methodenwünsche 8 (= logistische Regression) und 2 (= Regressionsanalyse) folgendermaßen mit dem sparsamen Set kategorialer Variablen METH1 bis METH3 erfasst:



Daraus ergeben sich folgende Werte für die Variablen STAT1 bis STAT8:



In obiger Syntax werden zwei ausgesprochen nützliche Kontrollstrukturen der SPSS-Kommandosprache verwendet:

Schleife für strukturgleiche Transformationen

Die (DO REPEAT - END REPEAT) - Schleife wird achtmal ausgeführt, wobei im *i*-ten Umlauf die beiden Stellvertreter STAT und N gerade mit den *i*-ten Elementen der zugehörigen Listen identisch sind.

Fallunterscheidung

Um in der folgenden Tabelle das SPSS-Verhalten beim Ausführen der (DO IF - ELSE - END IF) - Struktur in Abhängigkeit vom Wahrheitswert des logischen Ausdruck illustrieren zu können, versetzen wir uns in den Zustand vor der MD-Behandlung für die Variablen METH1, METH2 und METH3 zurück (vgl. Abschnitt 6.5.3):

Wert des logischen Ausdrucks	Aktion
wahr, z.B. im ersten Schleifenumlauf bei METH1 = 1, METH2 = 2, METH3 = SYSMIS	Das erste COMPUTE-Kommando wird ausgeführt.
falsch, z.B. im ersten Schleifenumlauf bei METH1 = 3, METH2 = 5, METH3 = 8	Das zweite COMPUTE-Kommando wird ausgeführt.
unbestimmt, z.B. im ersten Schleifenumlauf bei METH1=3, METH2=5, METH3=SYSMIS	Die neue Variable STAT1 behält den Initialisierungswert SYSMIS.

14 Datendateien im Textformat einlesen

Gelegentlich sind Daten auszuwerten, die in Textdateien vorliegen. In Abschnitt 3.1.2 wurden zwei Dateiformate beschrieben, die uns dabei begegnen können:

- positionierte Daten (feste Breite)
- separierte Daten (mit Trennzeichen).

Zum Importieren von Textdatendateien stellt SPSS einen leistungsfähigen Assistenten zur Verfügung, der mit

Datei > Textdaten lesen

gestartet wird. Er kommt aber auch dann zum Einsatz, wenn Sie nach

Datei > Öffnen > Daten

eine Textdatendatei wählen.

An der im Vorwort vereinbarten Stelle finden Sie die Dateien **kfar-kv-pos.txt** und **kfar-kv-sep.txt** mit positionierten bzw. separierten KFA-Daten von 77 Fällen. Es bietet sich an, diese Daten einzulesen, um die in Abschnitt 9.4 durch graphische Datenexploration gewonnene Moderatorversion der differentialpsychologischen Hypothese anhand einer unabhängigen Stichprobe zu überprüfen.

14.1 Import von positionierten Textdaten (feste Breite)

In der Datei **kfar-kv-pos.txt** sind die Werte eines Falles auf zwei Zeilen verteilt, und jede Variable hat eine feste Position im Datensatz eines Falles (z.B. Variable AERGO in Zeile 2, Spalten 5-6), so dass auch ihre Breite fixiert ist.

```
11 177115848
12  6 6 431214542432 110000
21 177115955
22  4 8 343335442442 110010
31 174416048
32  3 8 433224443342 100010
41 175116578
42  2 2 553125544531 100100
 .  .  .  .  .  .  .  .  .
 .  .  .  .  .  .  .  .  .
```

Die für uns relevanten Variablen haben folgende Positionen:

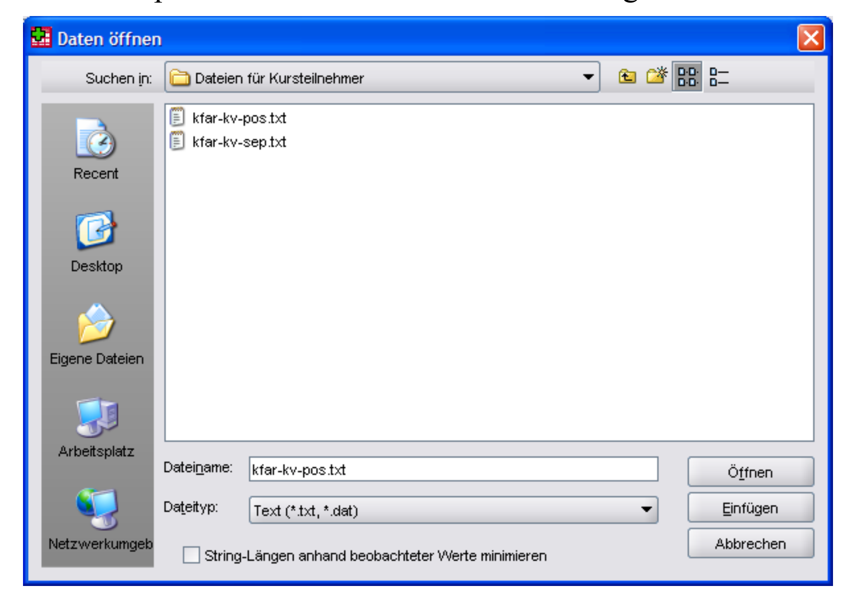
Variable	Datenzeile	Spalten
GESCHL	1	5
AERGO	2	5-6
AERGM	2	7-8
LOT01-LOT12	2	10-21

Alle übrigen Variablen können wir ignorieren.

Gehen Sie folgendermaßen vor, um die relevanten Daten zu importieren:

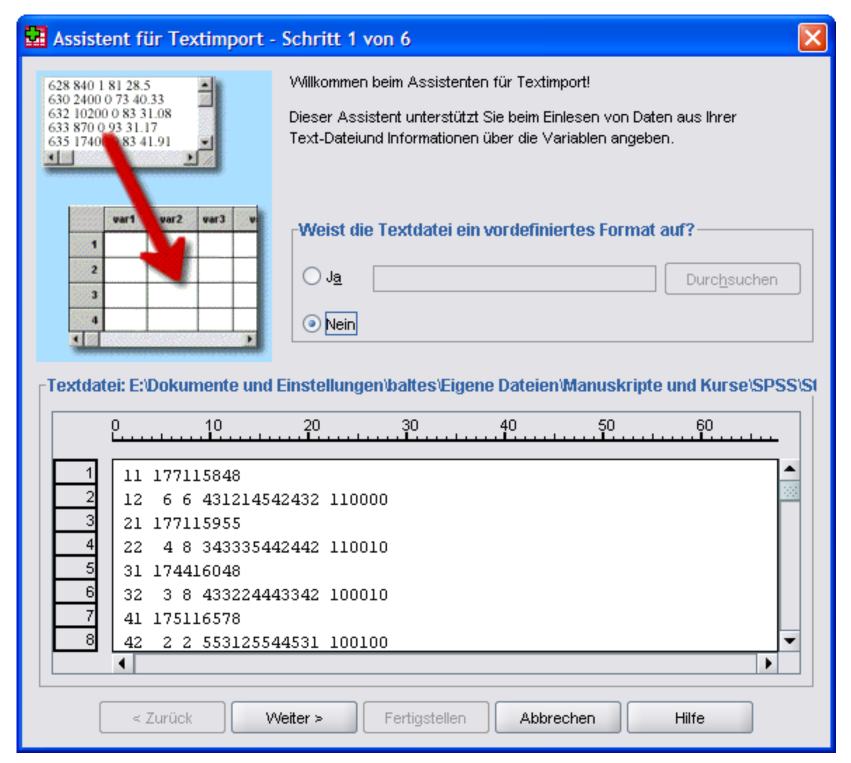
Textimport-Assistenten starten und Datei auswählen

Nach dem Start des Textimport-Assistenten ist zunächst die Eingabedatei zu wählen:



Schritt 1

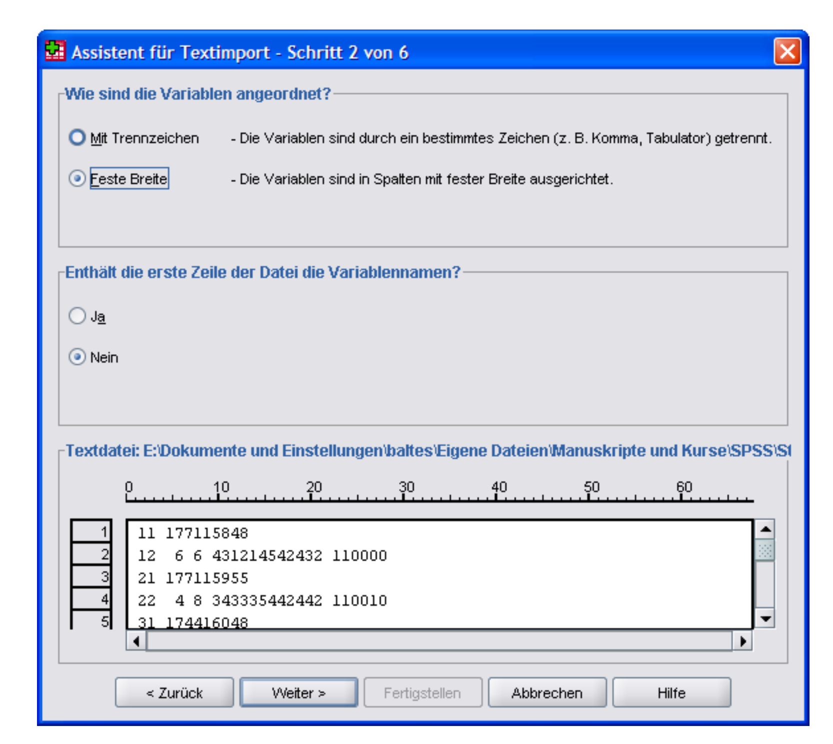
Im ersten Schritt zeigt der Assistent den Anfang unserer Datei und akzeptiert ggf. ein **vordefiniertes Format** aus früheren Assistenteneinsätzen, das die Dateistruktur beschreibt.



Da wir auf eine solche Vorarbeit *nicht* zurückgreifen können, machen wir **weiter**.

Schritt 2

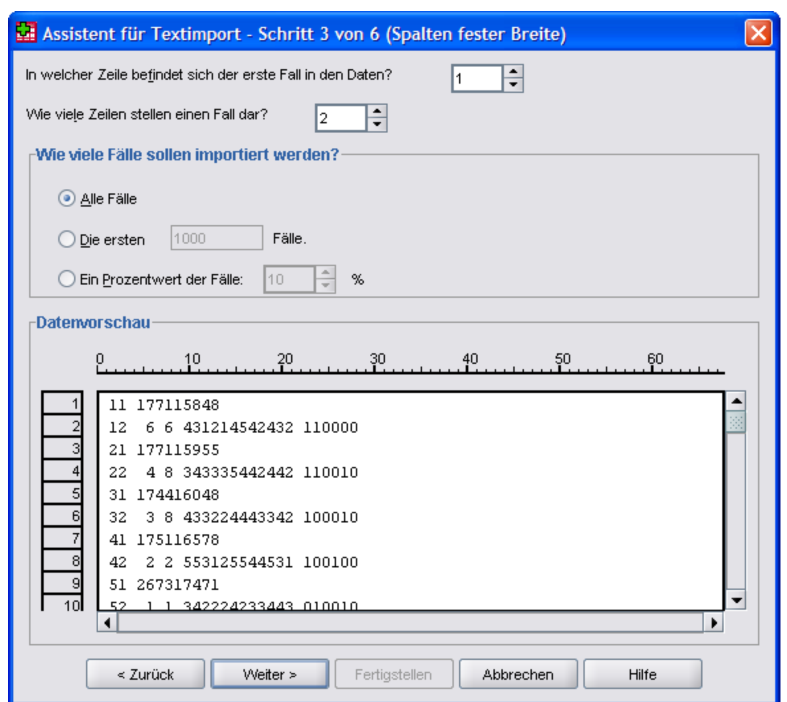
Im zweiten Schritt teilen wir mit, dass die Variablen in unserer Eingabedatei feste Positionen bzw. eine **feste Breite** besitzen:



Von der Möglichkeit, in der **ersten Zeile der Datei die Variablennamen** zu transportieren, wird in unserem Beispiel kein Gebrauch gemacht.

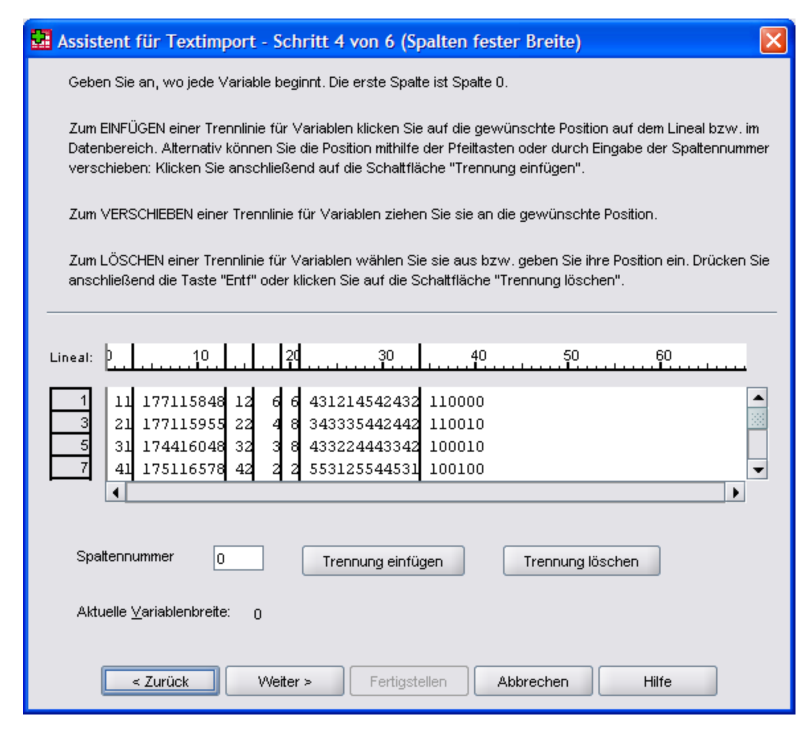
Schritt 3

Da unsere Datei keinen Vorspann enthält, **befindet sich der erste Fall** in Zeile 1. Allerdings befindet er sich dort nicht komplett, weil jeweils zwei **Zeilen einen Fall darstellen**:



Schritt 4

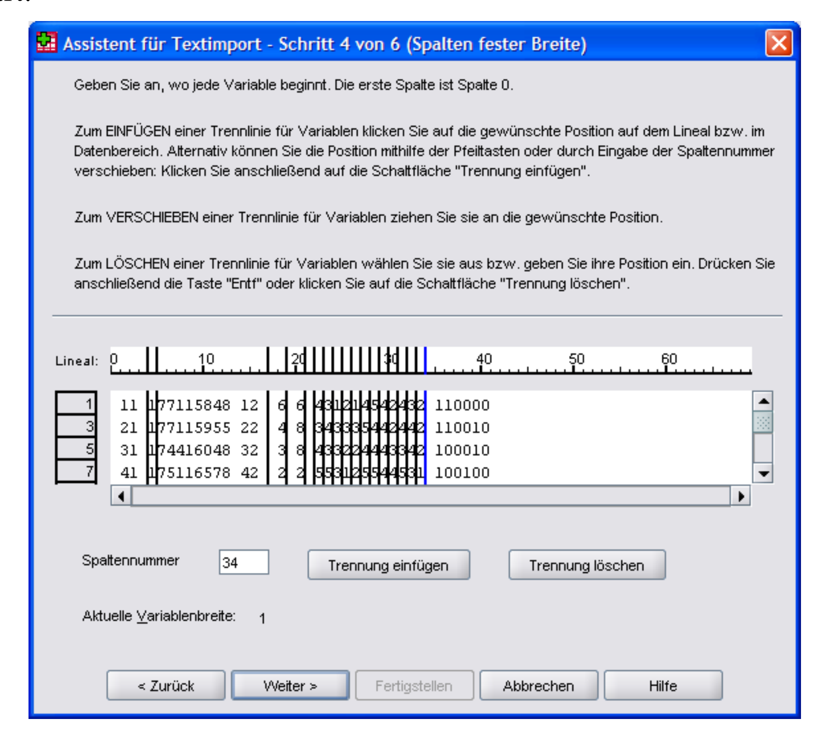
Nun müssen wir die Positionen der einzulesenden Variablen durch Setzen, Verschieben und Löschen von Trennlinien festlegen, wobei alle Zeilen eines Falles hintereinander angezeigt werden. Der Assistentenvorschlag orientiert sich an Leerzeichen und würde im Beispiel zu sieben, teilweise unbrauchbaren Variablen führen:



Hinweise zur Benutzung der Trennlinien:

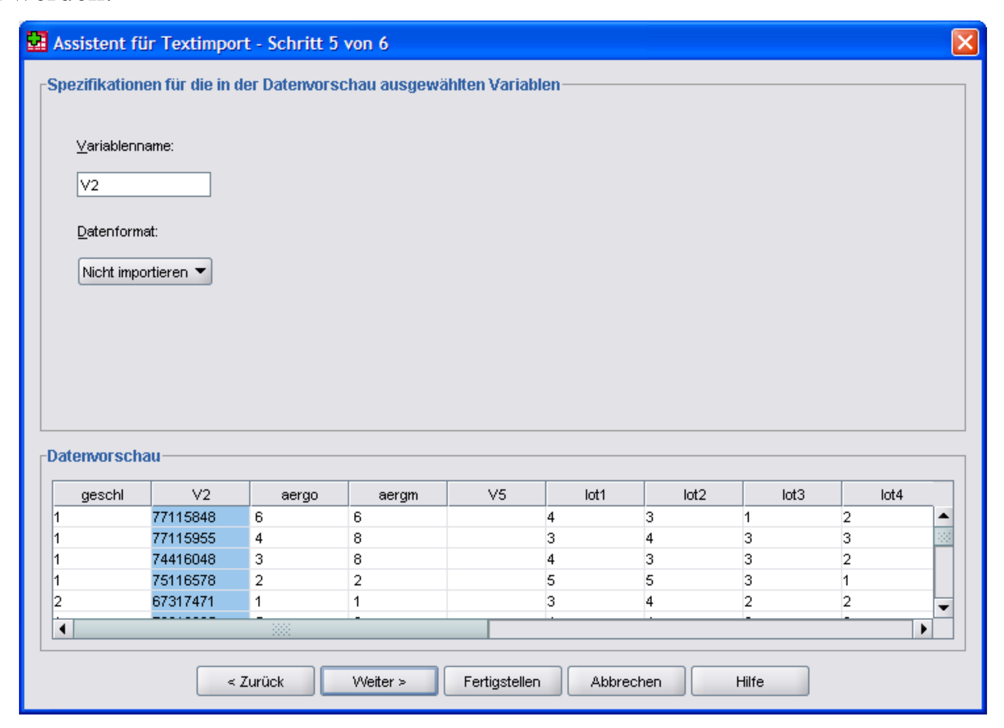
- Neue Trennlinie einfügen
  Klicken Sie innerhalb der Datenzone auf die gewünschte Spaltenposition.
- Trennlinie verschieben
  Klicken Sie innerhalb der Datenzone auf die Trennlinie und verschieben Sie diese bei fest gehaltener Maustaste.
- Trennlinie löschen
  Trennlinie per Mausklick markieren und anschließend **Trennung löschen**

So wird im Beispiel das Einlesen der Variablen GESCHL, AERGO, AERGM und LOT1 bis LOT12 vereinbart:



Schritt 5

Im fünften Assistentenschritt können wir die von SPSS vorgeschlagenen Variablennamen ändern und ein **Datenformat** festlegen. Zum Umbenennen ist jeweils genau eine Spalte zu markieren. Das Datenformat lässt sich auch für eine markierte Variablenliste wählen.
Mit dem speziellen Datenformat **Nicht importieren** können überflüssige Variablen ausgeschlossen werden:



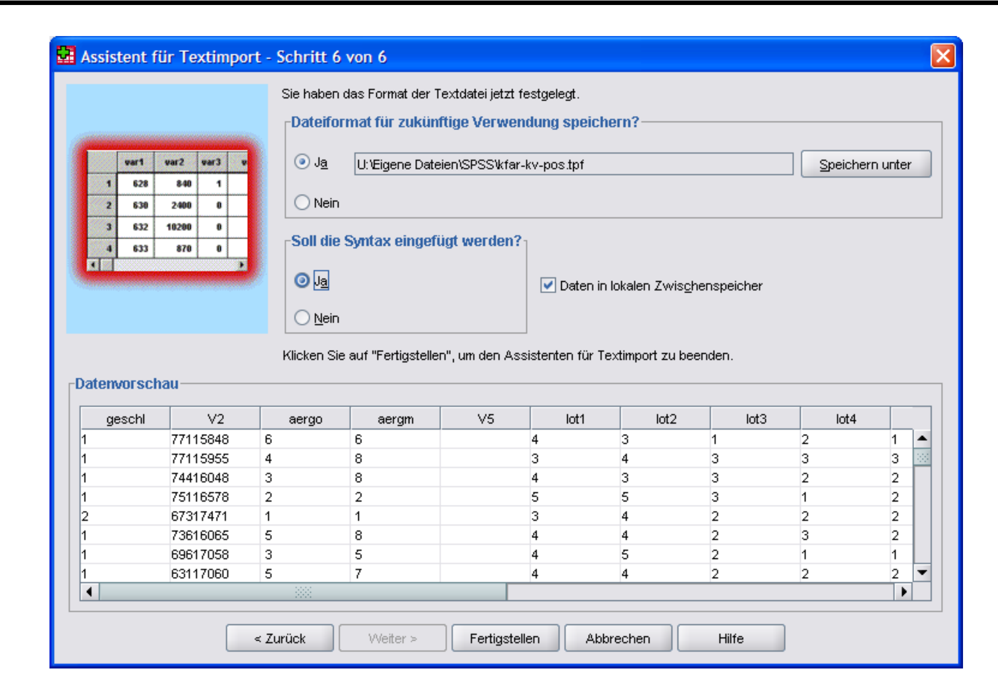
Zumindest bei den LOT-Variablen ist echte Fleißarbeit zu leisten, so dass wir nach Schritt 5 noch **weiter** machen, um unsere Arbeit zu konservieren.

Schritt 6

Der Assistent bietet zwei Möglichkeiten zum Konservieren einer Dateispezifikation:

- **Dateiformat für zukünftige Verwendung speichern?**
  Es entsteht eine Textassistenten-Formatdatei (Erweiterung **.tpf**), die bei einem späteren Assistenteneinsatz im ersten Schritt angegeben werden kann (siehe oben).
- **Soll die Syntax eingefügt werden?**
  Das für den Datenimport verantwortliche GET DATA – Kommando wird in ein Syntaxfenster geschrieben. Es bietet sich an, zusätzliche Kommandos zu ergänzen, z.B. zum Deklarieren von MD-Indikatoren, die in den Textdaten vorhanden sind. Später kann mit Hilfe des entstandenen SPSS-Programms der Import mit allen erforderlichen Zusatzmaßnahmen automatisiert ausgeführt werden.

Es spricht nichts dagegen, beide Konservierungsoptionen zu verwenden:



Das vom Textimport-Assistenten erzeugte GET DATA – Kommando verblüfft etwas mit einer Spaltenzählung ab **0**:

```
GET DATA
  /TYPE=TXT
  /FILE='U:\Eigene Dateien\SPSS\kfar-kv-pos.txt'
  /FIXCASE=2
  /ARRANGEMENT=FIXED
  /FIRSTCASE=1
  /IMPORTCASE=ALL
  /VARIABLES=
  /1 geschl 4-4 F1.0
  V2 5-12 8X
  /2 aergo 4-5 F2.0
  aergm 6-7 F2.0
  V5 8-8 1X
  lot1 9-9 F1.0
  lot2 10-10 F1.0
  lot3 11-11 F1.0
  lot4 12-12 F1.0
  lot5 13-13 F1.0
  lot6 14-14 F1.0
  lot7 15-15 F1.0
  lot8 16-16 F1.0
  lot9 17-17 F1.0
  lot10 18-18 F1.0
  lot11 19-19 F1.0
  lot12 20-20 F1.0
  V18 21-27 7X.
CACHE.
EXECUTE.
```

Nach dem Einlesen einer Textdatei dürfen Sie auf keinen Fall die Deklaration der dort eventuell verwendeten **MD-Indikatoren** vergessen. Studieren Sie also sorgfältig den hoffentlich vorhandenen Kodierplan, der in unserem Fall vorschreibt:

Variable	MD-Indikator
GESCHL	9
AERGO	99
AERGM	99
LOT1-LOT12	9

Die Deklaration kann in der Variablenansicht des Dateneditors erfolgen (siehe Abschnitt 3.2.2). Bei der Variablen AERGO ist z.B. für die Spalte **Fehlende Werte** einzutragen:



Das Kommando MISSING VALUES erlaubt allerdings eine rationellere (und automatisierbare) MD-Deklaration:

```
missing values  geschl (9) /aergo aergm (99) /lot1 to lot12 (9).
```

14.2 Import von separierten Daten Textdaten

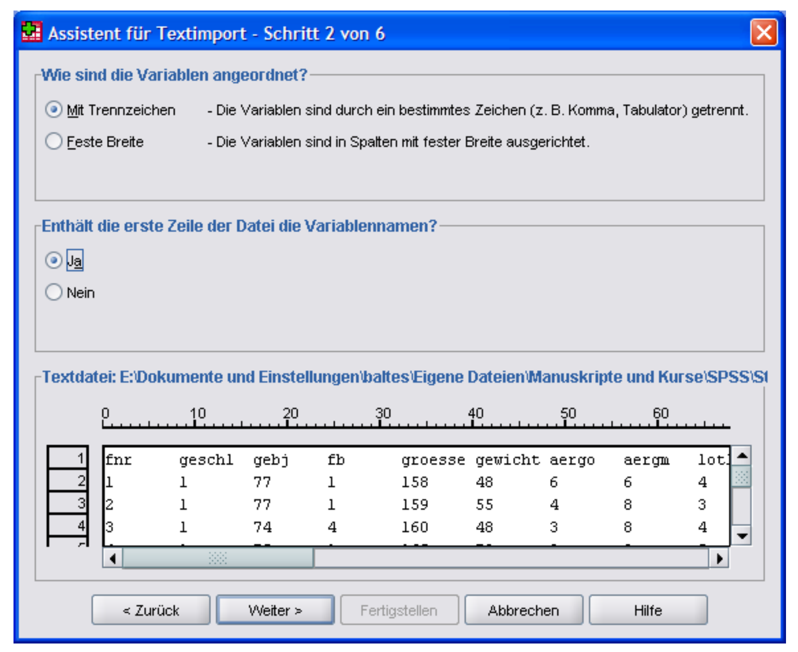
Separierte Textdaten lassen sich erheblich bequemer importieren als positionierte, zumal sie üblicherweise durch eine Zeile mit den Variablennamen eingeleitet werden. Die Datei **kfar-kv-sep.txt** enthält dieselben KFA-Daten, die in Abschnitt 14.1 aus einer positionierten Datei gelesen wurden:

FNR	GESCHL	GEBJ	FB	GROESSE	GEWICHT	AERGO	AERGM	LOT1	LOT2	...
1	1	77	1	158	48	6	6	4	3	...
2	1	77	1	159	55	4	8	3	4	...
3	1	74	4	160	48	3	8	4	3	...
4	1	75	1	165	78	2	2	5	5	...
.	.	.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	.	.	...

Beim Import der separierten KFA-Textdaten informieren wir den über

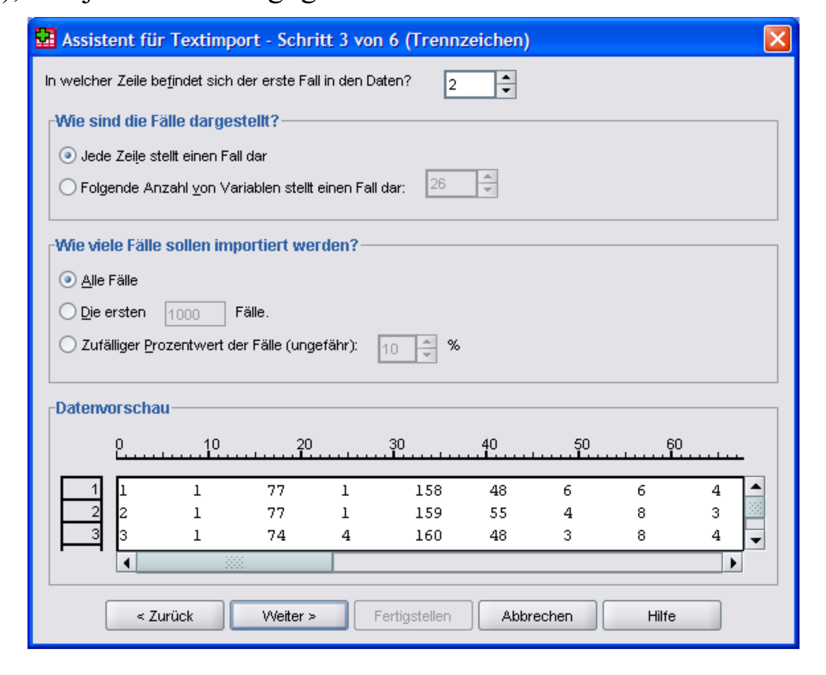
Datei > Textdaten lesen

gestarteten Assistenten im zweiten Schritt darüber, dass **Trennzeichen** für Ordnung in der Datei sorgen, und dass die erste Zeile die **Variablennamen** enthält:



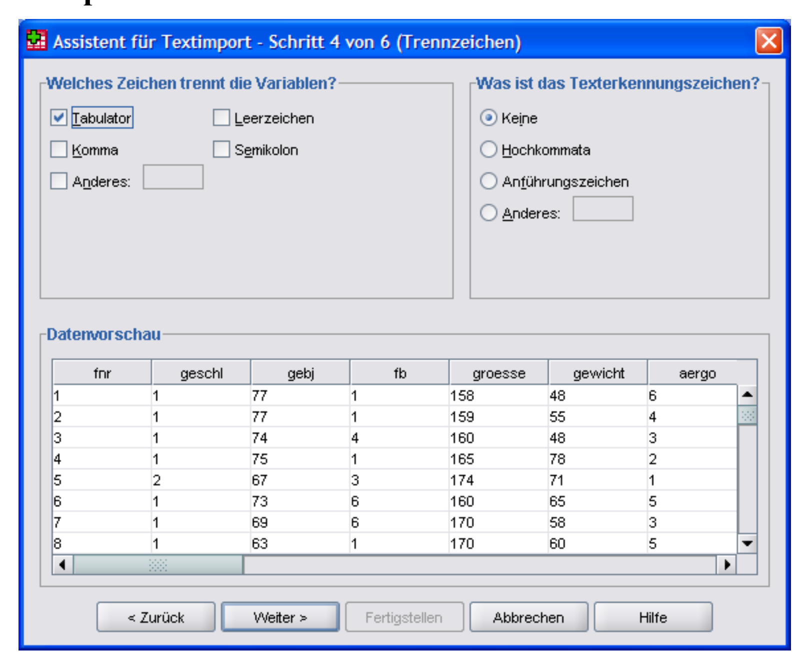
Schritt 3

Der erste Fall befindet sich in der zweiten Zeile der Datei (hinter der einleitenden Zeile mit den Variablennamen), und jeder Fall belegt genau eine Zeile:



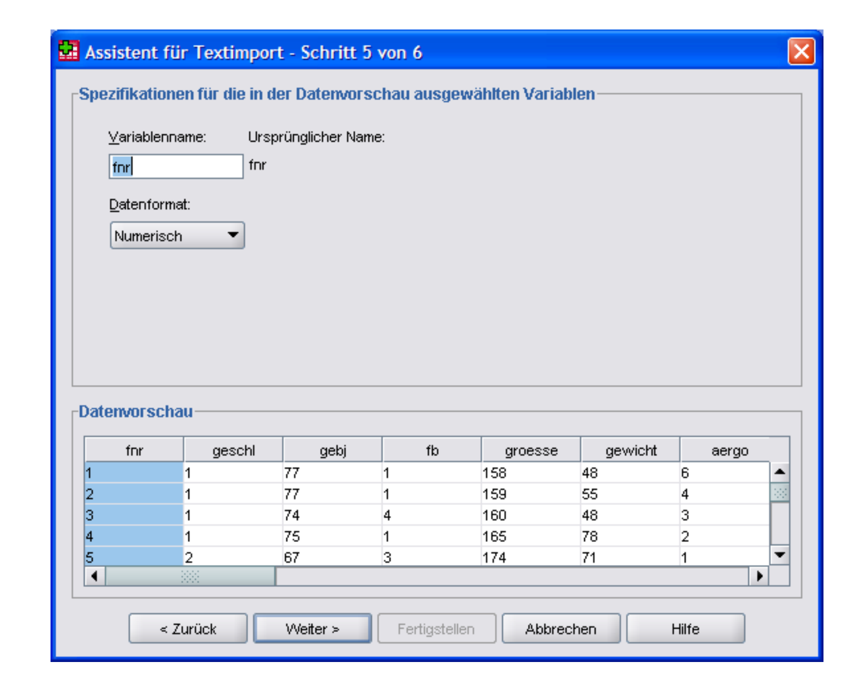
Schritt 4

In der Datei **kfar-kv-sep.txt** kommt als Trennzeichen nur der **Tabulator** zum Einsatz:



Schritt 5

Im fünften Assistentenschritt müssen wir nur prüfen, ob die automatische Erkennung des **Datenformats** erfolgreich war:



Schritt 6

Im letzten Assistentendialog werden die schon in Abschnitt 14.1 vorstellten Optionen zum Konservieren der Importspezifikation angeboten.

Auch nach dem Einlesen von separierten Textdaten dürfen Sie auf keinen Fall die Deklaration der eventuell vorhandenen **MD-Indikatoren** vergessen.

14.3 Überprüfung der revidierten differentialpsychologischen Hypothese

Um mit den in Abschnitt 14.1 bzw. Abschnitt 14.2 importierten Daten die revidierte differentialpsychologische Hypothese prüfen zu können, sind zunächst einige Datentransformationen erforderlich, wobei wir uns die erforderlichen Kommandos teilweise aus dem Transformationsprogramm **kfat.sps** besorgen können:

```
* Labels für GESCHL.
VARIABLE LABELS geschl Geschlecht.
VALUE LABELS geschl 1 'Frau' 2 'Mann'.

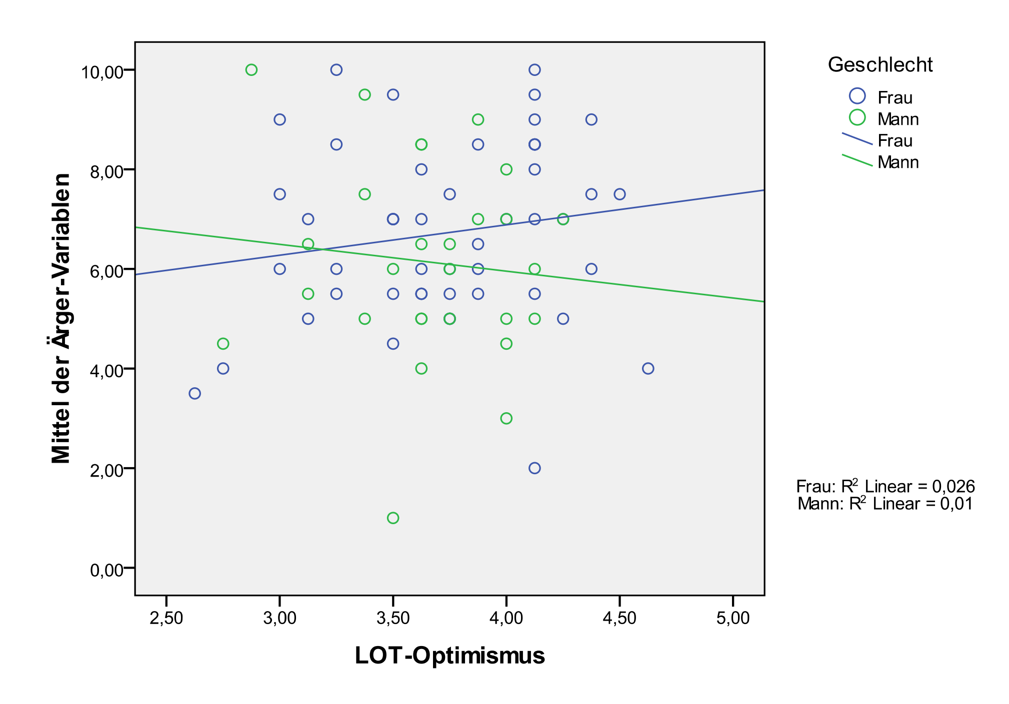
* LOT-Fragen umkodieren.
RECODE
  lot3 lot4 lot5 lot12  (5=1)  (4=2)  (2=4)  (1=5)  .
EXECUTE .

* LOT berechnen.
COMPUTE lot = MEAN.6(lot1,lot3,lot4,lot5,lot8,lot9,lot11,lot12) .
VARIABLE LABELS lot 'LOT-Optimismus' .
EXECUTE .

* AERGAM berechnen.
COMPUTE aergam = (aergo + aergm)/2 .
VARIABLE LABELS aergam 'Mittel der Ärger-Variablen' .
EXECUTE .

* Produktvariable für die Moderatorhypothese.
COMPUTE geslot = geschl * lot.
VARIABLE LABELS geslot 'GESCHL * LOT'.
EXECUTE .
```

Auch in der neuen Stichprobe scheint das Geschlecht die Regression von AERGAM auf LOT im erwarteten Sinn zu moderieren:



Allerdings wird der Interaktionseffekt in der Moderatoranalyse *nicht* signifikant (p = 0,307 im Test für den Produktterm):

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten: Regressions-koeffizient B	Nicht standardisierte Koeffizienten: Standardfehler	Standardisierte Koeffizienten: Beta	T	Sig.
1	(Konstante)	,773	5,562		,139	,890
	Geschlecht	3,670	4,130	,949	,889	,377
	LOT-Optimismus	1,761	1,493	,413	1,180	,242
	GESCHL * LOT	-1,150	1,118	-1,120	-1,029	,307

a. Abhängige Variable: Mittel der Ärger-Variablen

Weitere Versuche zur Rettung der differentialpsychologischen Hypothese könnten sich z.B. auf eventuelle Mängel bei der Operationalisierung der theoretischen Begriffe (Ärger und Optimismus) konzentrieren. Allerdings muss auch die theoretische Fundierung kritisch hinterfragt werden.

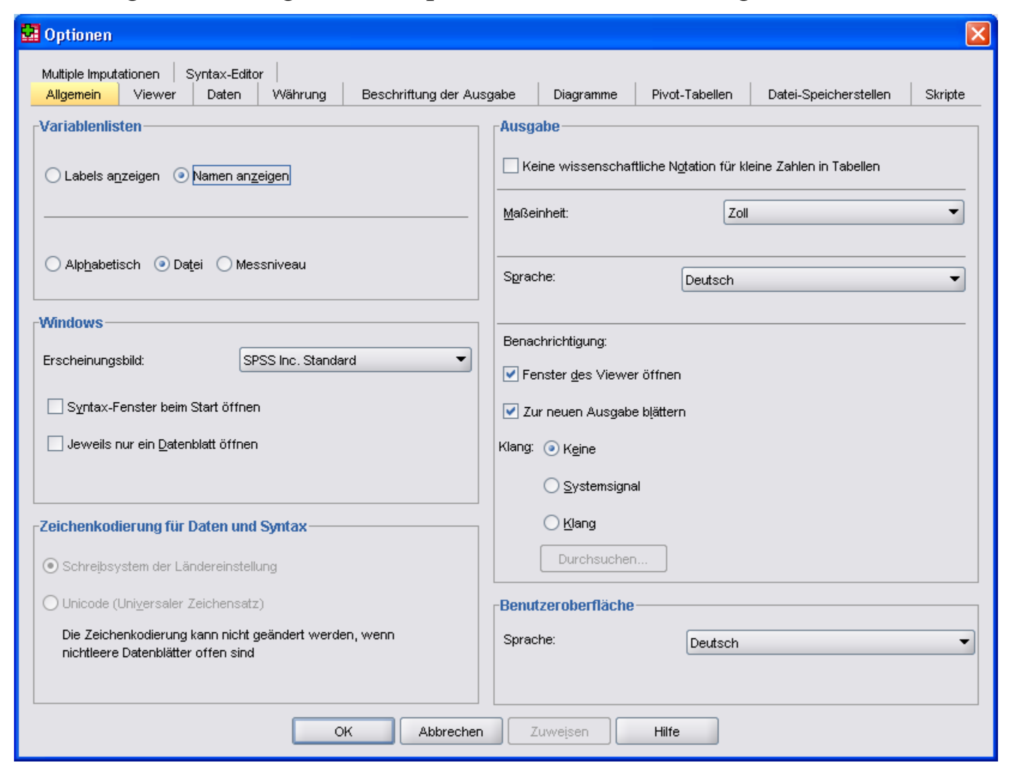
15 Einstellungen modifizieren

Das Standardverhalten von SPSS für Windows lässt sich auf vielfältige Weise den individuellen Bedürfnissen anpassen, was wir bei passender Gelegenheit auch schon getan haben.

Über den Menübefehl

Bearbeiten > Optionen

erhalten Sie die folgende Dialogbox mit Optionen zur SPSS-Konfiguration:



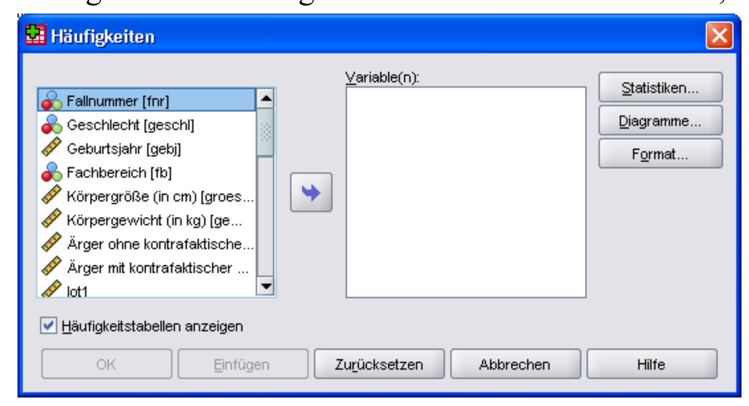
15.1 Allgemein

Auf dem Registerblatt **Allgemein** sind u.a. folgende Optionen von Relevanz:

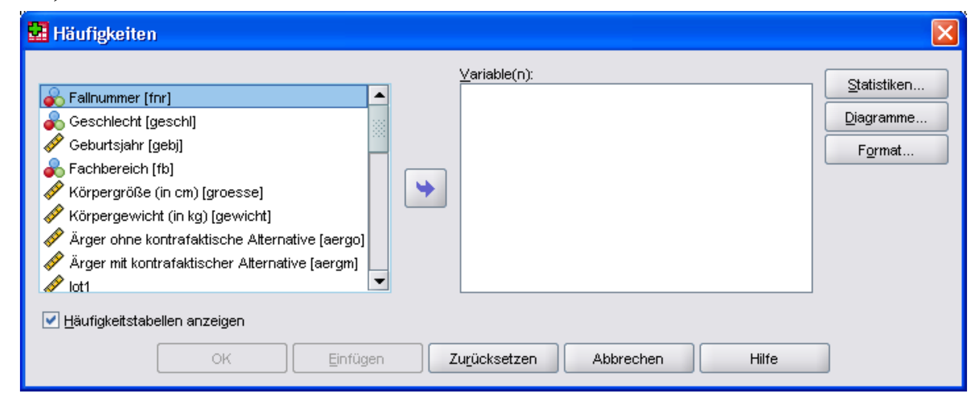
Variablenlisten

Bei den Listen auswählbarer Variablen in Dialogboxen verwendet SPSS folgende Voreinstellungen:

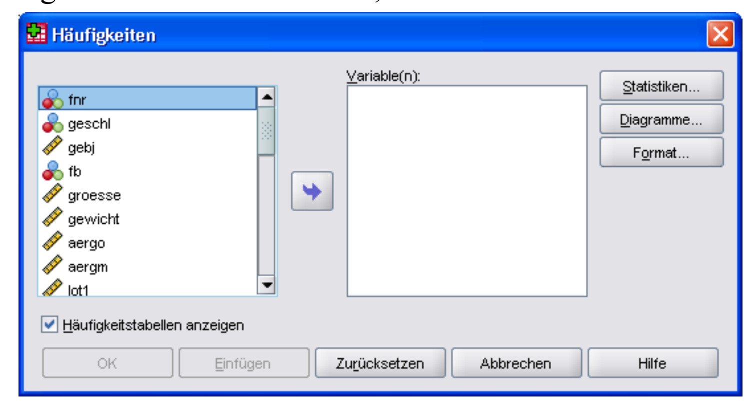
- Besitzt eine Variable ein Label, wird dieses vorrangig präsentiert, und der Variablenname erscheint hinter dem Label zwischen eckigen Klammern. Dabei werden die Variablenlisten aufgrund des begrenzten Platzangebotes oft recht unübersichtlich, z.B.:



Seit SPSS 16 kann man durch Verbreitern von Dialogboxen gekappte Variablenlabels vermeiden, z.B.:



Mit der Option **Namen anzeigen** im Bereich **Variablenlisten** kann man auf die kompaktere Darstellung *ohne* Labels umschalten, z.B.:



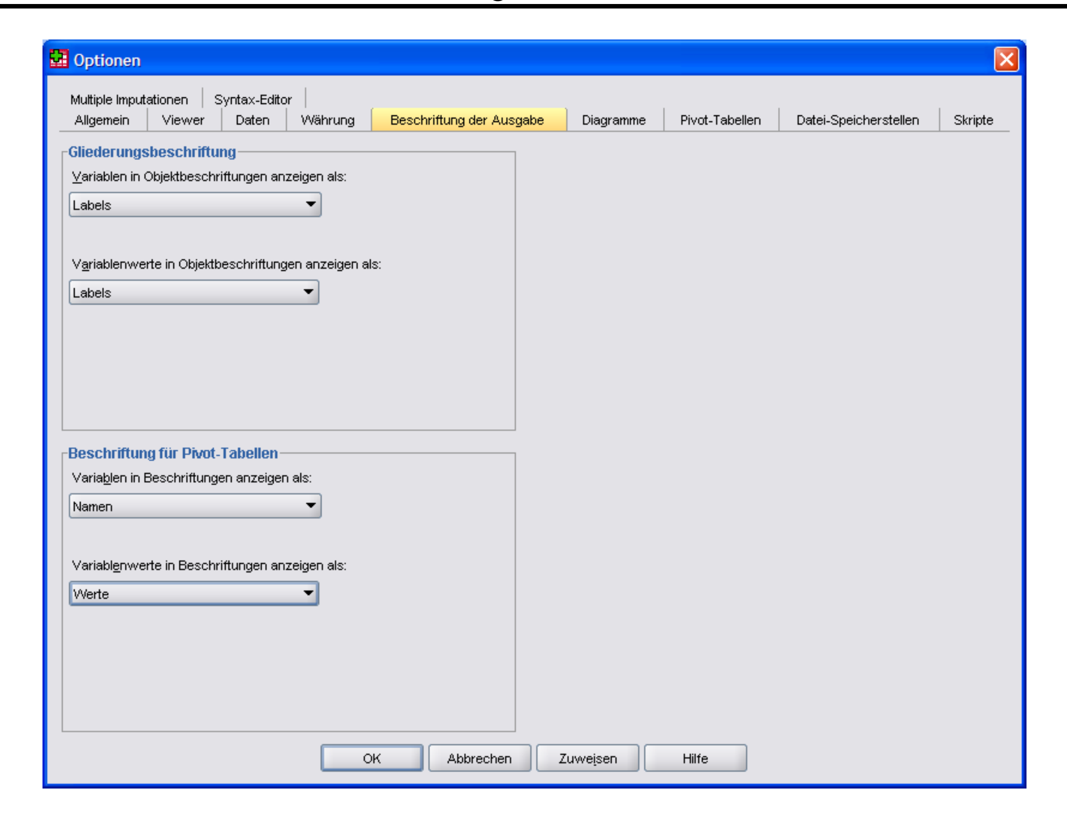
- Die Variablen sind angeordnet wie in der Arbeitsdatei, was in der Regel ein bequemes Arbeiten erlaubt. Gemeinsam zu analysierende und damit in Dialogboxen auszuwählende Variablen stehen nämlich oft in der Arbeitsdatei hintereinander. Bei der Arbeit mit einer unbekannten Datendatei findet man (namentlich bekannte) Variablen jedoch leichter bei alphanumerischer Sortierung.

Sprache der Ausgabe bzw. Benutzeroberfläche

SPSS erlaubt für die **Ausgabe** (Beschriftung der Tabellen) und für die **Benutzeroberfläche** (Menüs und Dialogboxen) die Wahl zwischen diversen Sprachen.

15.2 Beschriftung der Ausgabe

Auf dem Registerblatt **Beschriftung der Ausgabe** können Sie z.B. veranlassen, dass in Pivot-Tabellen vorhandene Wertelabels ignoriert und stattdessen die Werte selbst angezeigt werden:



15.3 Datei-Speicherstellen

Auf dem Registerblatt **Datei-Speicherstellen** kann man u.a. einstellen:

Startordner für die Dialogfelder zum Öffnen und Speichern

Auf den Pool-PCs an der Universität Trier eignet sich der Ordner

U:\Eigene Dateien\SPSS

Sitzungs-Journal

Per Voreinstellung protokolliert SPSS alle Kommandos, die Sie während einer Sitzung per Dialogbox oder via Syntaxfenster abschicken, in einer so genannten **Journaldatei**. Diese Datei kann für Anwender(innen) mit „Mut zur SPSS-Syntax“ sehr nützlich sein, weil sie die Kommando-Äquivalente zu praktisch allen Arbeiten früherer Sitzungen enthält. Per Voreinstellung wird beim Start einer SPSS-Sitzung eine vorhandene Journaldatei *nicht* überschrieben, sondern die neuen Kommandos werden am Ende angehängt. Falls die Datei zu groß wird, muss sie gelegentlich verkleinert oder gelöscht werden. Man kann aber auch im Rahmen **Sitzungs-Journal** den voreingestellten Öffnungsmodus **Anhängen** abändern auf **Überschreiben**. Dann wird die Journaldatei zu Beginn jeder Sitzung neu erstellt, wobei gegebenenfalls der alte Inhalt überschrieben wird.

Auf den Pool-PCs an der Universität Trier bietet sich die Verwendung der folgenden Datei an:

U:\Eigene Dateien\SPSS\spss.jnl

16 Anhang

16.1 *Weitere Hinweise zur SPSS-Kommandosprache*

In Abschnitt 5 wurden nur sehr oberflächliche Hinweise zur SPSS-Kommandosprache gegeben. Diese sollten genügen für Anwender(innen), die nicht frei programmieren, sondern nur gelegentlich ein von SPSS automatisch erzeugtes Kommando modifizieren wollen. Der aktuelle Abschnitt ist für ambitionierte Anwender(innen) gedacht, die bereit sind, SPSS-Programme zu schreiben, ...

- um auch die ausschließlich per Syntax verfügbaren SPSS-Leistungen nutzen zu können,
- um rationeller mit SPSS zu arbeiten.

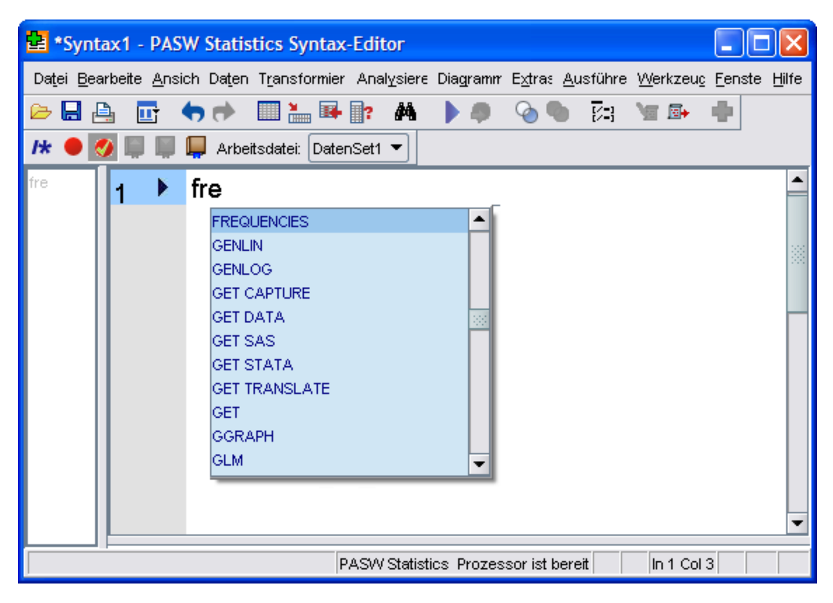
16.1.1 Hilfsmittel für das Arbeiten mit der SPSS-Kommandosprache

Das wichtigste Hilfsmittel für das Arbeiten mit der SPSS-Kommandosprache ist die *Command Syntax Reference*, die als PDF-Dokument über das Hilfesystem verfügbar ist:

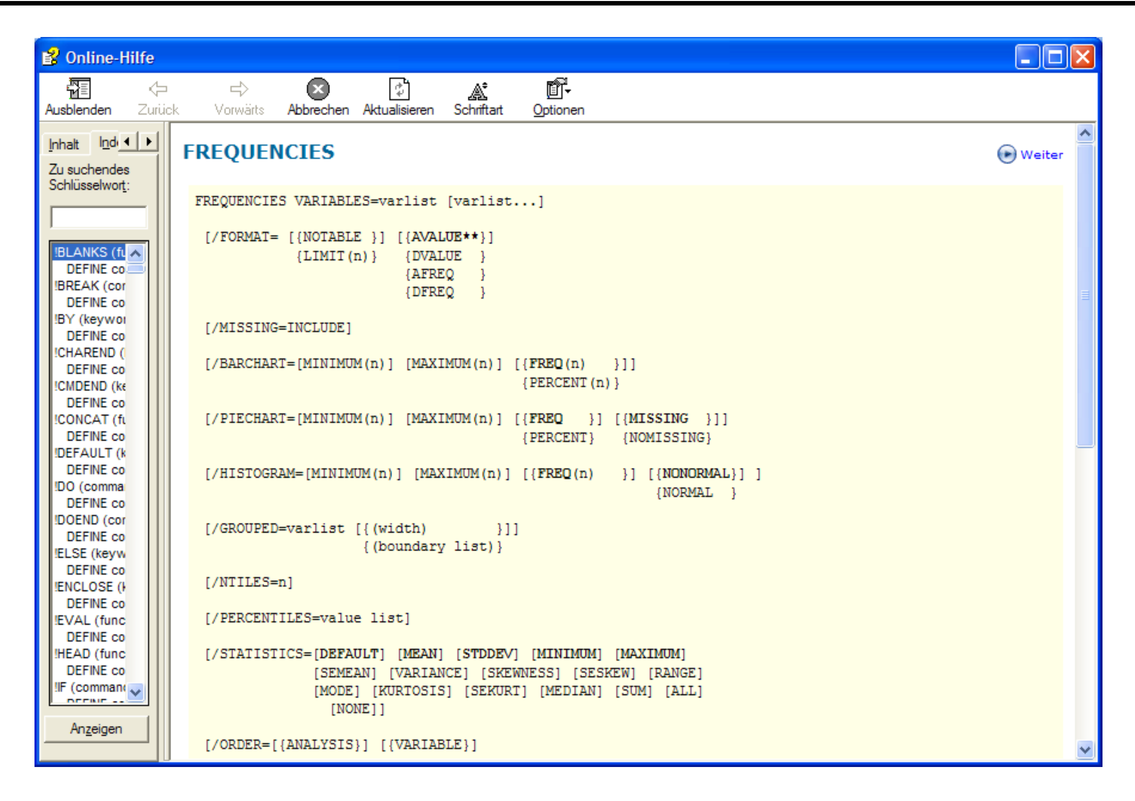
Hilfe > Befehlssyntax-Referenz

Hier findet man ausführliche Beschreibungen der SPSS-Kommandos mit zahlreichen Beispielen und wertvollen Literaturhinweisen zu den realisierten statistischen Methoden.

Das Syntaxfenster erleichtert die Bearbeitung von Kommandos über Zeilennummern, farbliche Unterscheidung verschiedener Syntaxbestandteile und eine intelligente Syntaxvervollständigung, z.B.:



Außerdem bietet es ein bequemes Verfahren, Syntaxinformationen zu einem konkreten Kommando anzufordern: Setzen Sie die Schreibmarke auf das Kommando, und klicken Sie dann auf das Symbol . Zum FREQUENCIES-Kommando, das der **Häufigkeiten**-Dialogbox zugrunde liegt, erscheint z.B. das folgende Hilfefenster:

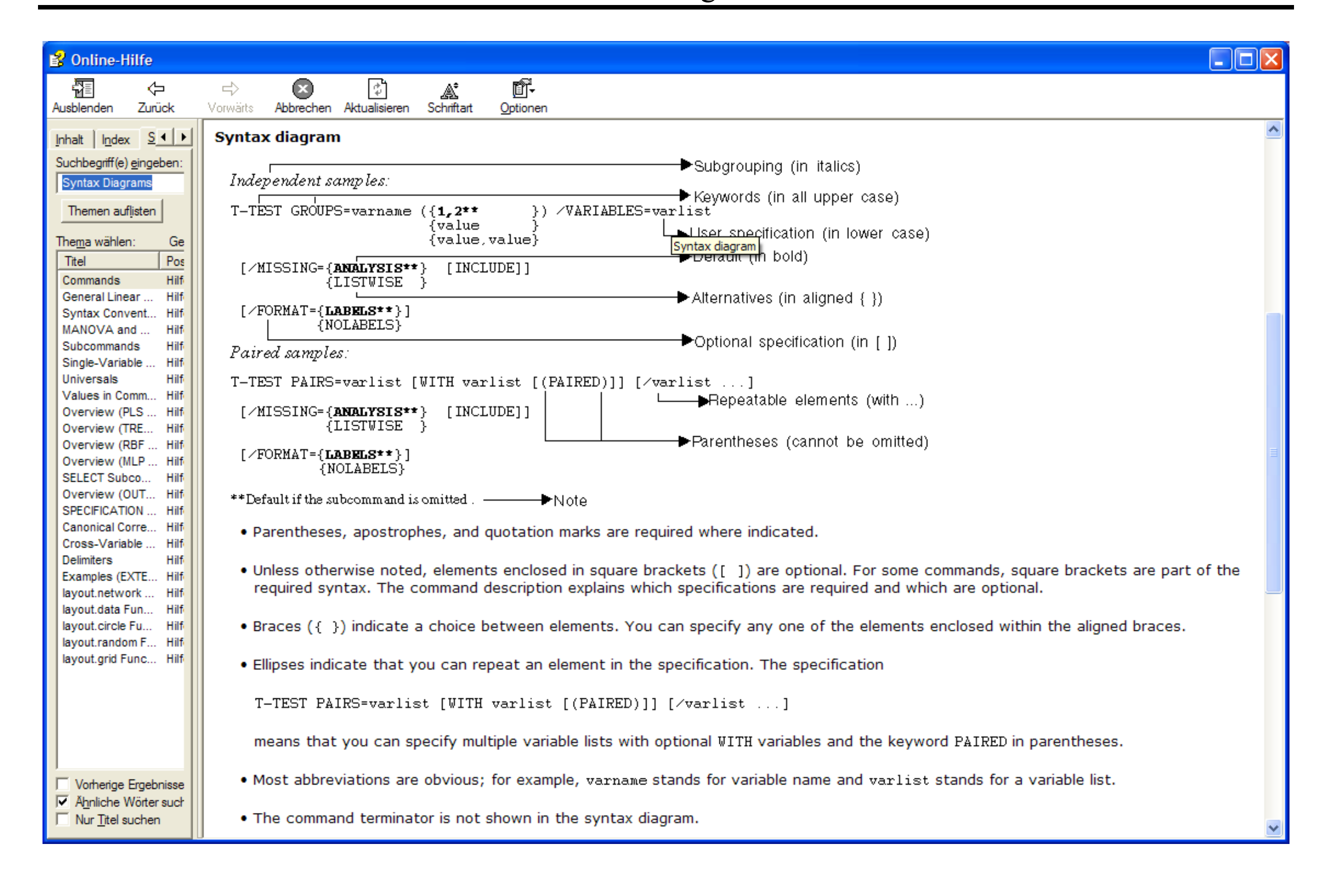


16.1.2 Interpretation von Syntaxdiagrammen

Mit einem Syntaxdiagramm wird die allgemeine Form eines Kommandos definiert und somit festgelegt, wie konkrete Beispiele gebildet werden müssen. Solche Syntaxdiagramme werden auch im weiteren Verlauf dieses Abschnitts benutzt, um Bestandteile der SPSS-Kommandosprache zu erläutern. In den Syntaxdiagrammen treten einige Metazeichen auf (z.B. "[", "{"), die nicht zur Kommandosprache selbst gehören, sondern diese Sprache beschreiben. Die Bedeutung dieser Metazeichen müssen Sie kennen, um Syntaxdiagramme richtig interpretieren zu können. Im Hilfesystem finden Sie eine Erläuterung, indem Sie nach

Hilfe > Themen > Suchen

den Suchbegriff *Syntax Diagrams* in das aktive Textfeld eintippen und dann einen Doppelklick auf das Thema **Commands** setzen:



16.1.3 Aufbau von SPSS-Programmen

Welche Kommandos SPSS für das Erstellen von Programmen bereithalten muss, ergibt sich aus unseren Zielvorstellungen: Wir möchten SPSS anweisen, unsere empirischen Daten zu lesen, gegebenenfalls aus den gelesenen Variablen interessante neue Variablen zu berechnen und schließlich statistische Verfahren mit den eingelesenen oder neu erstellten Variablen zu rechnen. Darüber hinaus haben wir gelegentlich Sonderwünsche hinsichtlich der Arbeitsweise von SPSS.

Orientiert an den gerade skizzierten Teilaufgaben kann man die verfügbaren SPSS-Kommandos in folgende Gruppen einteilen:

- **Dateidefinitions-Kommandos**
  Sie dienen zum Einlesen von Daten in die Arbeitsdatei. Als Beispiel haben wir bereits das GET-Kommando kennen gelernt. Wenn ein Programm kein Dateidefinitions-Kommando enthält, wenn es also nicht selbst für das Einlesen seiner Daten sorgt, kann es natürlich nur ausgeführt werden, wenn zuvor eine Arbeitsdatei erzeugt worden ist.
- **Transformations-Kommandos**
  Diese Kommandos dienen zur Veränderung oder Neuberechnung von Variablen bzw. zur Auswahl von Fällen für die weitere Verarbeitung.
- **Prozedur-Kommandos**
  Damit werden statistische Analysen, graphische Präsentationen oder Dateibearbeitungen (z.B. Sortieren der Fälle) angefordert. Ein Beispiel ist das FREQUENCIES-Kommando.
- **Dienst-Kommandos**
  Damit kann man u.a. die Arbeitsweise von SPSS beeinflussen (z.B. Startwert des Pseudozufallszahlengenerators setzen) und verschiedene Informationen anfordern.

In folgendem SPSS-Programm treten Kommandos aus allen Gruppen auf:

```
comment Größe und Gewicht.                      | Dienst-Kommando

get file = 'kfar.sav'.                          | Dateidef.-Kommando

frequencies var = groesse gewicht               | Prozedur-
 /statistics = all /histogram = normal.         | Kommando

compute ideal = groesse - 100.                  | Transformations-
                                                | Kommando

t-test pairs = gewicht ideal.                   | Prozedur-
                                                | Kommando
```

SPSS-Programme können flexibel gestaltet werden:

- In einem Programm dürfen beliebig viele Prozedur-Kommandos auftreten.
  Manche Anwender leben in dem Irrglauben, pro SPSS-Programm sei nur eine einzige Statistikprozedur erlaubt, und verstreuen daher zusammenhängende Auswertungen über unübersichtlich viele Miniprogramme. Andere haben den falschen Ehrgeiz, ihr gesamtes Projekt in einem einzigen Programm abzuwickeln, und erstellen dabei ein unpraktisches Monsterprogramm mit mehreren hundert Zeilen. Wie so oft im Leben ist auch hier der gesunde Mittelweg zu empfehlen: Für abgrenzbare Aufgabenpakete sollte jeweils ein eigenes Programm erstellt werden (z.B. mit allen Prozeduren zur Datenprüfung).
- Auch *nach* einer Prozedur dürfen Datentransformationen vorgenommen werden.
- Man kann nach einer Prozedur sogar weitermachen mit der Definition einer neuen Arbeitsdatei, welche dann die alte ersetzt.

16.1.4 Aufbau eines einzelnen SPSS-Kommandos

Die wichtigsten Regeln für SPSS-Befehle:

- Ein Kommando besteht aus seinem Namen und den zugehörigen Spezifikationen:

kommandoname spezifikationen

  - Der **Kommandoname** kann aus einem Wort bestehen oder aus mehreren Wörtern.
    Beispiele:
    - FREQUENCIES
    - GET DATA
  - Die **Spezifikationen** dürfen enthalten:
    - Schlüsselwörter (z.B. VARIABLES)
    - Variablennamen
    - Zahlen
    - Zeichenfolgen (z.B. Variablenlabel)
    - Operatoren (z.B. "+")
    - spezielle Begrenzungszeichen: / ( ) = ' "

    Zwischen diesen Elementen ist mindestens ein Leerzeichen erforderlich. Ausnahme:
    Die speziellen Begrenzungszeichen, die arithmetischen Operatoren und manche Vergleichsoperatoren (z.B. ">") sind selbstbegrenzend, d.h. davor und danach sind keine Leerzeichen nötig (aber erlaubt).

Statt eines Leerzeichens darf man meist verwenden:
- beliebig viele Leerzeichen,
- ein Komma,
- einen Zeilenwechsel.

Dies ermöglicht eine übersichtliche Programmgestaltung.

- Jedes Kommando muss in einer neuen Zeile beginnen und mit einem Punkt enden.
  Die Kommandos müssen dabei keinesfalls in der ersten *Spalte* beginnen, sondern dürfen eingerückt werden. Von dieser Möglichkeit sollte man z.B. bei Schleifen-Konstruktionen Gebrauch machen.
  Beispiel:
  ```
  do repeat  mc=mc001 to mc100.
    compute  mc=normal(1).
  end repeat.
  ```
  Hier werden 100 unabhängige, normalverteilte Zufallsvariablen erzeugt. Durch das Einrücken wird deutlich gemacht, dass die COMPUTE-Anweisung innerhalb der DO REPEAT - Schleife steht.

- Ein Kommando kann sich über beliebig viele Fortsetzungszeilen erstrecken. *Innerhalb* eines Kommandos sind aber keine Leerzeilen erlaubt.

- Eine Syntaxzeile sollte maximal 256 Zeichen enthalten, um in allen Kontexten ausführbar zu sein.

- Die Verwendung von Groß- oder Kleinbuchstaben ist beliebig.

- Schlüsselwörter dürfen meist bis auf die ersten drei Zeichen abgekürzt werden.
  Beispiel: `"fre"` für `"frequencies"`

- Bei den meisten Kommandos sind die Spezifikationen in Subkommandos unterteilt. Diese beginnen mit einem Subkommando-Namen, meist gefolgt von einem Gleichheitszeichen, und sind durch Schrägstriche voneinander getrennt.
  Beispiel:
  ```
  frequencies var=lot01 /format=notable
    /statistics=all.
  ```

Merken Sie sich aus dieser Liste für den Anfang vor allem:

JEDES KOMMANDO MUSS IN EINER NEUEN ZEILE BEGINNEN UND MIT EINEM PUNKT ENDEN.

16.1.5 Regeln für Variablenlisten

16.1.5.1 Abkürzende Spezifikation einer Serie von Variablen

In Transformations- oder Prozedurkommandos soll häufig eine Folge **bereits existierender** und **in der Arbeitsdatei hintereinander liegender** Variablen angesprochen werden. Dies ermöglicht das **aufrufende TO**, dessen Syntax im Folgenden erläutert wird:

vara TO *varb*

vara, *varb*	Namen bereits vorhandener Variablen, wobei *vara* in der Arbeitsdatei vor *varb* stehen muss.

Beispiele: - `frequencies var=alter to beruf.`
Für alle Variablen, die in der Arbeitsdatei von ALTER bis BERUF positioniert sind, werden Häufigkeitstabellen erstellt.
- `frequencies var=frage1 to frage3.`
Wenn in der Arbeitsdatei zwischen FRAGE1 und FRAGE3 1500 beliebig benannte Variablen stehen, dann bewirkt dieses Kommando 1502 Häufigkeitstabellen!

16.1.5.2 Der Platzhalter varlist

In folgendem Syntaxdiagramm wird der in SPSS-Kommandos häufig auftretende Platzhalter *varlist* definiert:

{*varname* | *varname_1* TO *varname_2*} [{...]

varname, *varname_1*, *varname_2* Variablennamen

Beispiel: `missing values  nieder01 to hoehe ozon mess1 to mess4 (9).`
Hier wird mit dem MISSING VALUES - Kommando für alle aufgelisteten Variablen die 9 als MD-Indikator vereinbart.

Literaturverzeichnis

Backhaus, K., Erichson, B, Plinke, W. & Weiber, R. (2008). *Multivariate Analysemethoden* (12. Aufl.). Berlin: Springer.

Baltes-Götz, B. (1998). *Exakte Tests mit SPSS*. Online-Dokumentation: http://www.uni-trier.de/index.php?id=22571

Baltes-Götz, B. (2008a). *Lineare Regressionsanalyse mit SPSS*. Online-Dokumentation: http://www.uni-trier.de/index.php?id=22489

Baltes-Götz, B. (2008b). *Behandlung fehlender Werte in SPSS und Amos*. Online-Dokumentation: http://www.uni-trier.de/index.php?id=23239

Baltes-Götz, B. (2008c). *Logistische Regressionsanalyse mit SPSS*. Online-Dokumentation: http://www.uni-trier.de/index.php?id=22513

Baltes-Götz, B. (2009). *Moderatoranalyse per multipler Regression mit SPSS*. Online-Dokumentation: http://www.uni-trier.de/index.php?id=22528

Bortz, J. (1977). *Lehrbuch der Statistik*. Berlin: Springer.

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation*. Berlin: Springer.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences.* New York: Academic Press.

Cohen, J., Cohen, P., West, S.G. & Aiken, L. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). Mahwah: Lawrence Erlbaum Associates.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.

Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149-1160.

Field, A. (2005). *Discovering Statistics Using SPSS*. London: Sage.

Hartung, J. (1989). *Statistik* (7. Auflage). München: Oldenbourg.

Kahneman, D. & Miller, D.T. (1986) Norm theory: comparing reality to its alternatives. *Psychological Review*, 93, 136-153.

Mehta, C.R., Patel, N.R. (1996). *SPSS Exact Tests 7.0 for Windows*. Chicago, IL: SPSS Inc.

Norušis, M.J. (2009). *SPSS 16.0. Statistical Procedures Companion*. Upper Saddle River, NJ: Prentice Hall.

Norušis, M.J. (2009). *SPSS 16.0. Advanced Statistical Procedures Companion*. Upper Saddle River, NJ: Prentice Hall.

Pedhazur, E.J. & Pedhazur Schmelkin L. (1991). *Measurement, design, and analysis. An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models* (2nd ed.). Thousand Oaks, CA: Sage.

Scheier, M.F. & Carver, C.S. (1985). Optimism, Coping, Health: Assessment and implications of generalized outcome expectancies. *Health Psychology*, 4, 219-247.

Schnell, R., Hill, P. B. & Esser, E. (2005). *Methoden der empirischen Sozialforschung* (7. Aufl.). München: Oldenbourg.

Siegel, S. (1976). *Nichtparametrische statistische Methoden*. Frankfurt: Fachbuchhandlung für Psychologie

Tabachnik, B.G. & Fidell, L.S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson.

Stevens, J. (1996). Applied *Multivariate Statistics for the Social Sciences* (3rd ed.). Mahwah: Lawrence Erlbaum.

Wallis, W.A. & Roberts, H.V. (1956). *Statistics, a new approach*. Glencoe, Ill.: The Free Press.

Wentura, D. (2004). Ein kleiner Leitfaden zur Teststärke-Analyse. Online-Dokument: http://www.uni-saarland.de/fak5/excops/download/POWER.pdf

Stichwortregister

F

G

H

I

J

K

L

M

N

O

P

R

S